МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

С. В. Горобець, О. Ю. Горобець, М. О. Булаєвська

БІОІНФОРМАТИЧНІ БАЗИ ДАНИХ

Рекомендовано Методичною радою КПІ ім. Ігоря Сікорського
як навчальний посібник для студентів,
які навчаються за спеціальністю 162 «Біотехнології та біоінженерія»

Київ
КПІ ім. Ігоря Сікорського
2020

Рецензент *Галкін О.Ю.*, д.б.н, професор, завідувач кафедри трансляційної медичної біоінженерії факультету біомедичної інженерії Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського»

Відповідальний редактор *Горго Ю.П.*, д.б.н, професор

Гриф надано Методичною радою КПІ ім. Ігоря Сікорського (протокол № 10 від 18.06.2020 р.) за поданням Вченої ради Факультету біотехнології і біотехніки (протокол № 10 від 25.05.2020 р.)

Електронне мережне навчальне видання

Горобець Світлана Василівна, д-р техн. наук, проф.
Горобець Оксана Юріївна, д-р. фіз.-мат. наук, проф.
Булаєвська Марина Олександрівна

БІОІНФОРМАТИЧНІ БАЗИ ДАНИХ

Біоінформатичні бази даних: [Електронний ресурс] : навч. посіб. для студ. спеціальності 162 «Біотехнології та біоінженерія» / С. В. Горобець, О. Ю. Горобець, М. О. Булаєвська; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 3,86 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2020. – 117 с.

В навчальному посібнику «Біоінформатичні бази даних» висвітлені призначення та класифікація баз даних, методики пошуку інформації в них, опис найбільш популярних біоінформатичних баз даних. Біоінформатичні бази даних призначені для зберігання і систематизації амінокислотних і нуклеотидних послідовностей різних організмів та їх порівняльного аналізу з метою розв'язання задач геноміки, протеоміки, метаболоміки, теорії еволюції, медицини, генної інженерії тощо. Робота з базами даних та використання методів біоінформатики є новим інструментом в молекулярній біології, біотехнології, медицині для отримання нових знань, що стоїть в одному ряду з фізичними та біохімічними методами досліджень.

Посібник рекомендується до використання в освітній діяльності для забезпечення підготовки бакалаврів та магістрів спеціальності 162-Біотехнології та біоінженерія.

© С. В. Горобець, О. Ю. Горобець, М. О. Булаєвська, 2020
© КПІ ім. Ігоря Сікорського, 2020

Зміст

СПИСОК СКОРОЧЕНЬ

AsMamDB – Alternative Splice Data Base of Mammal; база даних альтернативного сплайсингу у ссавців

BCR – **Biospecimen Core Resource**; централізований сайт, що послідовно оцінює патологію та дані ДНК та РНК

BiGG – Biochamical Genetic and Genomic data base; база даних метаболізму людини

BioCyc – **BioCyc is a collection of 17043 Pathway/Genome Databases (PGDBs))** – колекція із 371 бази даних геномів/метаболічних шляхів

BLAST – Basic Local Alignment Search Tool; засіб пошуку основного локального вирівнювання

BLASTn – вирівнювання нуклеотидних послідовностей

BLASTp – вирівнювання амінокислотних послідовностей

BLASTx – вирівнювання всіх можливих транслятів нашої нуклеотидної послідовності проти банка амінокислотних послідовностей

BMRB – The Biological Magnetic Resonance Data Bank; банк даних магнітного резонансу біооб’єктів

CCLE – Cancer Cell Line Encyclopedia**;** енциклопедія ракових клітин

ChEBI – Chemical Entities of Biological Interest; словник молекулярних об’єктів, орієнтований на невеликі хімічні сполуки

ChemSpider – database of chemicals**;** база даних хімічних сполук і сумішей

COSMIC – Catalogue of Somatic Mutations in Cancer; каталог соматичних мутацій пухлин людини

dbSNP – Single Nucleotide Polymorphism Database; база даних по однонуклеотидним поліморфізмам

DDBJ – DNA Data Bank of Japan; Японська база даних ДНК

DrugBank – база даних лікарських речовин з хімічної, фармакологічної і фармацевтичної інформацією

EBI – European Bioinformatics Institute; Європейський інститут біоінформатики

Entrez – пошукова система баз даних NCBI

EMBL – European Molecular Biology Laboratory; Європейська лабораторія молекулярної біології

EST – short sub-sequence of a cDNA sequence – короткі суб-послідовності ДНК

FDA – Food and Drug Administration; Управління нагляду за якістю харчових продуктів та медикаментів

FlyBase – онлайн-база даних біоінформатики та основне сховище генетичних та молекулярних даних для сімейства комах Drosophilidae.

FooDB – The Food Database; база даних продуктів харчування

GC – Gas chromatography; газова хроматографія

GC-MS – газова хроматографія-мас-спектроскопія

GDC – Genomic Data Commons; база даних досліджень раку

GenBank – база даних послідовностей ДНК

GO – Gene Ontology; генетична онтологія

GWAS – Genome-Wide Association Studies; повногеномний пошук асоціацій

HMDB – Human Metabolome Database; база даних метаболому людини

HPRD – Human Protein Reference Database; довідкова база даних білків людини

HumanCyc –база даних, що містить інформацію про метаболічні шляхи та геном людини

IIMDB – *In Vivo/In Silico* Metabolites Database; база даних, що містить як відомі, так і розраховані сполуки

INSDC – International Nucleotide Sequence Database Collaboration; міжнародна система баз даних нуклеотидних послідовностей

International HapMap Project – організація, метою якої є розвиток карти гаплотипів людського генома, яка описує загальні патерни генетичної мінливості у людей

KEGG – Kyoto Encyclopedia of Genes and Genomes; Кіотська енциклопедія генів і геномів

KEGG Glycan – база даних містить набір експериментально визначених структур гліканів великої кількості еукаріотичних та прокаріотичних організмів

LC – Liquid chromatography; рідинна хроматографія

LIMS – Laboratory Information Management System; система управління лабораторною інформацією, програмне забезпечення, призначене для управління потоками лабораторними робіт і документів

MassBank – мас-спектральна база даних експериментально отриманих мас-спектрів метаболітів

MetaboLights – база даних експериментів метаболоміки та отриманої інформації

MetaCyc – база даних не надлишкових експериментально з'ясованих метаболічних шляхів

METAGENE – база даних з інформацією про вроджені помилки метаболізму.

Metlin – найбільше сховище мас-спектральних даних метаболітів

MMDB – Molecular Modelling Database, база даних по молекулярному моделюванню

MMCD – Madison Metabolomics Consortium Database; база даних малих молекул, що зібрана з електронних баз даних та наукової літератури

MS – Mass-spectroscopy; мас-спектроскопія

NAR – журнал Nucleic Acid Research

NCBI – National Center for Biotechnology Information; Національний центр біотехнологічної інформації

NCI – National Cancer Institute; Національний інститут раку

NHGRI – National Institute for Human Genome Research; Національний інститут досліджень геному людини

OMIM – Online Mendelian Inheritance in Man; всебічний збірник генів людини та генетичних фенотипів

OMMBID – On-Line Metabolic and Molecular Basis to Inherited Disease інтернет-доступна енциклопедія, що описує генетику, метаболізм, діагностику та лікування сотень порушень обміну речовин

PDB – Protein Database, банк даних білків

PGP – Personal Genom Project; Персональний геномний проект

PSD-PIR – Protein Sequence Database-Protein Information Resource; база даних білків-білковий інформаційний ресурс

ProDom – Protein Domain, база даних білкових доменів

PROSITE – Protein Domain Database or Functional Characterization and Annotation; база даних білкових доменів для функціональної характеристики та анотації

PubChem – Public database of Chemical molecules**;** база даних хімічних структур малих органічних молекул та інформацією про їх біологічну активність

PubMed – бази даних статей в галузі медицини та біології

RCSB – Research Collaboratory for Structural Bioinformatics; міжнародна організація, яка забезпечує науково-дослідну співпрацю зі структурної біоінформатики

Reactome – database of reactions, pathways and biological processes; база даних біологічних шляхів

SAAS – System of Automatic Anotation of Statistics**;** система автоматизованої анотації статистичних даних

SCOP – бази даних структурної класифікації білків

SIB – Swiss Institute of Bioinformatics; Швейцарський інститут біоінформатики

SNP – Single nucleotide polymorphism; однонуклеотидний поліморфізм

SNPedia – сайт біоінформатики на основі вікі, який служить базою даних однонуклеотидних поліморфізмів

TaxBrowser – Taxonomic Browser; база даних таксономічної інформації

TBLASTx – програма для вирівнювання всіх можливих транслятів досліджуваної нуклеотидної послідовності проти всіх транслятів банка нуклеотидних послідовностей

TCGA – The Cancer Genome Atlas; атлас генома раку

TrEMBL – Translated EMBL**;** база даних амінокислотних послідовностей білків, що заповнюється шляхом автоматизованої вибірки даних з EMBL

UCSC – University of California Santa Cruz; Каліфорнійський університет Санта-Крус

UniParc – UniProt Archive; архів бази даних UniProt

UniProt – Universal Protein Resource; база даних послідовностей білків

UniProtKB – UniProt Knowledge base; база знань UniProt

UniRef – UniProt Reference Clusters; база даних посилань UniProt

WikiPathways – відкрита платформа для збору та розповсюдження моделей біологічних шляхів для візуалізації та аналізу даних

БД – бази даних

ФРТ – фактор росту тромбоцитів

ЯМР – ядерний магнітний резонанс

Розділ 1. ОСНОВИ БІОІНФОРМАТИЧНИХ БАЗ ДАНИХ

У 1965 р. американським вченим з фізичної хімії, піонером в області біоінформатики Маргарет Дейхофф зі співробітниками Національного Фонду біомедичних досліджень (National Medical Research Foundation, Вашингтон) було систематизовано усі наявні дані про амінокислотні послідовності і створено першу біоінформатичну базу даних – атлас білкових послідовностей та їх структур. Перша версія атласу містила опис 65 послідовностей білків. З тих пір науковцями світу було накопичено колосальний експериментальний матеріал про будову і функціонування біологічних молекул (ДНК, РНК, білків) (табл. 1.1). Цей матеріал для свого аналізу потребує розвинутих комп'ютерних методів, що привело до становлення та розвитку нового наукового напрямку – біоінформатики та створення біоінформатичних баз даних білків і нуклеотидних послідовностей.

Таблиця 1.1 – Розміри геномів деяких організмів

Таксон	**Розмір геному (середній по таксону, пар основ)**
Мікоплазми	$1{,}62 \cdot 10^6$
Бактерії	$2 \cdot 10^6$
Гриби	$2 \cdot 10^7$
Neurospora crassa	$4{,}7 \cdot 10^7$
Дріжджі	$1{,}4 \cdot 10^7$
Нематода *Caenorhabditis elegans*	$1 \cdot 10^8$
Комахи	$2{,}3 \cdot 10^9$
Дрозофіла	$1{,}8 \cdot 10^8$
Тутовий шовкопряд	$5 \cdot 10^8$

Кінець таблиці 1.1

Таксон	**Розмір геному (середній по таксону, пар основ)**
Молюски	$1{,}6 \cdot 10^9$
Костисті риби	$1{,}4 \cdot 10^9$
Хвостаті амфібії	$3{,}6 \cdot 10^{10}$
Безхвості амфібії	$2{,}7 \cdot 10^9$
Рептилії	$1{,}5 \cdot 10^9$
Птахи	$1{,}2 \cdot 10^9$
Ссавці	$2{,}6 \cdot 10^9$
Домова миша	$3 \cdot 10^9$
Людина	$3 \cdot 10^9$
Голонасінні	$1{,}6 \cdot 10^{10}$
Покритонасінні	$2{,}7 \cdot 10^{10}$
Кукурудза	$8 \cdot 10^9$

Кожна розшифрована нуклеотидна або амінокислотна послідовність сама по собі представляє значний інтерес для генної інженерії та біотехнології, але разом з цим, послідовність може слугувати колосальним джерелом інформації при порівнянні її з іншими послідовностями. Це стимулювало розробку добре структурованих баз даних (БД), які могли б працювати з великими об’ємами експериментальних даних, і спеціальних програмних засобів, що допомогли б інтерпретувати результати експериментів. Відомості про відкриття нових послідовностей розміщувалися у провідних журналах, а потім ці дані заносилися до БД вручну. Однак, коли почалося лавиноподібне зростання обсягів інформації, такий процес став неможливим. Журнали почали вимагати, щоб послідовності розміщувалися у БД самими авторами. Сьогодні, коли секвенування ДНК є достатньо поширеним процесом, який можуть виконувати роботи або студенти на лабораторних роботах, багато

послідовностей можуть потрапляти до БД без опублікування у наукових журналах. На даний час існують такі БД, де інформацію може розмістити кожен науковець, і такі, де інформація суворо перевіряється, а відповідальність за її достовірність покладається на власника бази даних. Так, база даних **GenBank** у Центрі біотехнологічної інформації США (National Center for Biotechnology Information (**NCBI)**), база даних EMBL-EBI у європейській лабораторії молекулярної біології (European Molecular Biology Laboratory) і Європейському інституті біоінформатики (European Bioinformatics Institute (EBI) та японська база даних **DDBJ** постійно обмінюються інформацією, містять всю відому інформацію про геноми організмів, заноситься інформація в ці БД самими науковцями.

Оволодіння навичками користування інтернет-ресурсами молекулярної біології відкриває широкі можливості у використанні біоінформатики (або обчислювальної молекулярної біології) не лише для пошуку і аналізу вже існуючої інформації, але й для отримання нових знань з меншими затратами матеріальних та часових ресурсів порівняно з фізико-хімічними дослідженнями. В багатьох випадках біоінформатичний аналіз з використанням геномних БД дозволяє отримати нові, нетривіальні висновки, тобто нові знання, які за необхідності потім можуть бути перевірені експериментально.

Вважається, що першим важливим з біологічної точки зору результатом, який було отримано за допомогою аналізу послідовностей, тобто методів біоінформатики, було виявлення подібності вірусного онкогену V-sis і нормального гену фактора росту тромбоцитів, що сприяло значному прогресу у розумінні механізму раку. Це відкриття подібних послідовностей з використанням алгоритму **локального вирівнювання** відбулося в 1983 році (Р. Дулітл і М. Уотерфільд). Вірусний онкоген мавпи V-sis виявився видозмінною формою нормального клітинного гену, який кодує фактор росту тромбоцитів (104 а.к.)

Фактор росту тромбоцитів (ФРТ) на 87% виявився подібним до білку вірусного онкогену V-sis, який може при визначених умовах запускати безконтрольну проліферацію клітин, перетворивши їх в злоякісні. ФРТ – це головний білок сироватки крові, необхідний для росту нормальних клітин в культурі тканин. Пухлинні ж клітини пред'являють знижені вимоги до факторів росту. Це спостереження давно викликало припущення про те, що пухлинні клітини здатні виробляти свій фактор росту, природа якого, однак, залишалася невідомою. У всіх клітинах хребетних, включаючи клітини людини, знаходиться ген, що кодує ФРТ, але **працює ФРТ тільки в тромбоцитах** (кров'яних пластинках), в інших же клітинах ФРТ не активний, «мовчить». Активною в інших клітинах може бути мутантна формо ФРТ, яким може бути, наприклад, вірусний онкоген V-sis. Тромбоцити мають обмежений термін життя. З їх загибеллю припиняється вироблення фактора росту. Але якщо той же ген починає працювати в клітинах шкіри, м'язовій або в клітинах будь-яких інших тканин, ситуація змінюється: клітина буде нестримно рости і ділитися. Встановлюється певний цикл, який і веде до розвитку пухлини. Незвичайна робота гену і синтез ФРТ іншими клітинами може бути пов'язана з мутацією, викликаною радіацією, хімічною речовиною, порушенням контролю транскрипції гена або трансляції мРНК. Оскільки ген, що кодує ФРТ, схожий з протоонкогеном V-sis, то йому може належати роль запуску складного багатоступінчастого пухлинного процесу. Зрозуміло, тромбоцитарний фактор росту впливає лише на клітини, які мають до нього рецептори. Такі клітини мезенхімального походження (в цьому секрет тканинно-специфічної дії онкогену). Всі ці дані спонукали дослідників до пошуків ФРТ в пухлинах, причому саме сполучних тканин, які мають до ФРТ рецептори. Виявилося, що 8 з 11 ліній пухлинних клітин, що ведуть своє походження з кісткової і сполучної тканин, виявили активне продукування фактора росту, схожого з ФРТ. Він був відкритий в остеосаркомі людини, в клітинах, трансформованих мавпячим ДНК-вірусом. Це підтвердило думку

про роль ФРТ в канцерогенезі. Відтоді робота з послідовностями стала необхідним елементом біологічних досліджень.

Тому біоінформатика є новим інструментом в біології, що стоїть в одному ряду з фізичними (рентгеноструктурний аналіз, електронна та скануюча мікроскопії, мас-спектрометрія, ядерний магнітний резонанс та ін.) та біохімічними методами досліджень.

Таким чином, окрім високого професійного рівня в області біології, ключовим моментом в оволодінні всіма можливостями сучасних біоінформатичних технологій є наявність у фахівця навичок швидкого та ефективного пошуку інформації в спеціалізованих БД через мережу Інтернет. В цьому аспекті слід згадати журнал Nucleic Acid Research (NAR), який з 1980 року перший номер року присвячує опису біоінформатичних БД, а з 2007 року замість цього випускає спеціальний додатковий номер.

1.1 Призначення та класифікація баз даних

З 1982 р. по молекулярній біології, зокрема, нуклеотидним та амінокислотним послідовностям, створено декілька тисяч БД. Серед них є електронні бази даних як загального призначення, так і вузькоспеціалізовані, а також БД наукових робіт з біології та медицини (Medline/PubMed та інші). Взагалі, **база даних** – це сукупність пов’язаної інформації, що об’єднана за певними ознаками. Більшість БД для збереження даних використовують таблиці. Кожна таблиця складається з рядків та стовпчиків, які називаються записами та полями, відповідно. Один запис може містити багато однакових полів, в які заноситься різна інформація.

БД оперують з наступними об’єктами:

- таблиці для збереження даних, що складаються зі строк та стовпців, або інакше, з записів та полів;
- запити для пошуку та одержання тільки необхідних даних;

- форми для перегляду, додавання та зміни даних в таблицях;
- звіти для аналізу та виводу даних в заданому форматі;
- сторінки доступу до даних через Інтернет.

Біоінформатичні БД забезпечують зручне та ефективне зберігання великої кількості інформації, систематизацію амінокислотних і нуклеотидних послідовностей, спрямовану на їх порівняльний аналіз, зокрема для:

- транслювання амінокислотних послідовностей білків;
- ідентифікації організмів, їх таксономічної приналежності і рівнів еволюційного розвитку, побудови філогенетичних дерев;
- задач генної інженерії;
- виявлення у неперервній послідовності символів окремих структурних одиниць та визначення їх функціонального навантаження;
- розшифровки просторової структури білків;
- виявлення структурно-функціональних взаємозв'язків груп білків;
- виявлення генів, які кодують макромолекули – потенційні мішені дії нових ліків та їх синтез (drug-desing).

Основним призначенням БД, крім збереження інформації, є швидкий пошук та цілеспрямоване структурування останньої. Електронні БД також повинні забезпечувати користувача засобами для управління всіма їх даними та інструментами для аналізу відповідної інформації.

Біоінформатичні БД призначені для зберігання і систематизації амінокислотних і нуклеотидних послідовностей різних організмів та їх порівняльного аналізу з метою розв'язання задач:

геноміки

- функціональна анотація окремих генів і повних геномів,
- передбачення по послідовностям просторової структури біополімерів,
- виявлення структурно-функціональних взаємозв'язків груп білків,

протеоміки (аналіз білків на рівні цілого генома),

метаболоміки (аналіз метаболізму шляхом одночасного вимірювання концентрацій багатьох речовин в клітинах, метаболічна реконструкція, аналіз регуляторних систем, тощо),

теорії еволюції

- від еволюції окремих генів і білків до еволюції метаболічних шляхів, регуляторних систем і цілих геномів,
- реконструкція початкових етапів виникнення генетичної інформації,
- ідентифікація організмів, їх таксономічної приналежності та ступенів еволюційного розвитку, побудова філогенетичних дерев,

медицини (виявлення генів, що є потенційними мішенями дії нових ліків та синтез останніх),

генної інженерії (цілеспрямована зміна генетичної інформації).

Існує багато різних типів баз даних, які відрізняються за джерелом надходження інформації, за тематикою тощо. Розглянемо класифікацію молекулярно-біологічних БД за джерелом надходження інформації:

Архівні БД – найбільш об'ємні та найменш достовірні, поповнюються самими дослідниками через Інтернет, а також з наукової літератури. Архівні, або первинні БД містять необроблені дані у тій формі, у якій вони були отримані із джерел. До них відносяться: **EMBL** (Європа), **GenBank** (США), **DDBJ** (Японія) – найбільші представники БД послідовностей нуклеотидів, **PDB** – БД просторових структур білків та багато інших БД.

БД, що куруються – це вторинні БД, які містять відібрану після аналізу з архівних БД інформацію. Інформацію з архівних БД і/або наукових публікацій відбирають експерти, перевіряючи її достовірність: **SwissProt** – найбільш якісна БД амінокислотних послідовностей білків, **KEGG** – інформація про метаболізм (для метаболічних шляхів), **COG –** інформація про ортологічні гени, **FlyBase** – інформація про *Drosophila* та ін.

Похідні БД та інтегровані БД – є результатом обробки даних з архівних БД і БД, що куруються. Якщо у таку БД ввести назву гену можна знайти усе, що про нього відомо – у яких організмах він зустрічається, у якому місці він

локалізований, які функції виконує, яку просторову структуру він має і таке інше. Це такі БД як, наприклад, БД **TrEMBL –** БД амінокислотних послідовностей білків, що заповнюється шляхом автоматизованої вибірки необхідних даних з БД EMBL. До похідних БД відноситься багато спеціалізованих БД: **SCOP –** БД структурної класифікації білків; **PFAM –** БД по сімействам білків; **GO (Gene Ontology) –** класифікація генів (термінологія); **ProDom –** БД білкових доменів; **AsMamDB –** БД альтернативного сплайсингу у ссавців; **SPN –** БД однонуклеотидних поліморфізмів; **Ecocyc –** БД бактерії *E. coli* (гени, білки, метаболізм та ін.).

Інші БД – по науковій літературі, по спеціалізованому програмному забезпеченню, по анатомії, біохімії та ін. Існують навіть БД по базам даних.

Записи БД містять нові експериментальні результати і додаткові відомості у формі анотацій. Анотації дають інформацію про джерела даних і методи отримання цих даних. Також надається інформація про дослідників і перелік публікацій по даному питанню. Не менш важливим є те, що записи забезпечують посилання на відповідні записи інших БД.

1.2 Методики пошуку інформації у БД

Усі існуючі бази даних надають можливісь роботи з ними через Internet та практично усі вони використовують стандартні методики пошуку, наприклад, можливість роботи з пошуковими системами:

Entrez (пошук по назві, номеру, організму, автору тощо). Забезпечує доступ до амінокислотних і нуклеотидних послідовностей, їх тривимірних структур, а також до повних секвенованих геномів, надає графічне відображення генів. Практично для кожної послідовності можна підібрати подібні послідовності та вже розраховані і визначені дво- та тривимірні структури, що відносяться до даної послідовності.

BLAST (basic local alignment search tool, пошук за подібністю) – порівнює надану інформацію з послідовностями, що вже є в базі для пошуку подібних

послідовностей. Є різні модифікації програми BLAST: BLASTp (вирівнювання амінокислотних послідовностей), BLASTn (вирівнювання нуклеотидних послідовностей), BLASTx (вирівнювання всіх можливих транслятів досліджуваної нуклеотидної послідовності з амінокислотними послідовностями БД), TBLASTx (вирівнювання всіх можливих транслятів досліджуваної нуклеотидної послідовності з всіма транслятами БД нуклеотидних послідовностей).

Offline-інтерфейс – спочатку з мережі Інтернет на локальний комп'ютер скачується частина бази даних, потім з цією частиною проводиться подальша робота.

Режим клієнт-сервер – на локальному комп'ютері встановлюється програма математичної обробки нуклеотидних послідовностей або послідовностей амінокислот, далі дана програма з'єднується з сервером бази даних і обробляє інформацію без скачування останньої на локальний комп'ютер. В той же час інтенсивно розвиваються системи обробки інформації та пошукові системи, що збирають і обробляють інформацію відповідно до запитів користувачів.

Програмне забезпечення баз даних повинно задовільняти наступним функціональним вимогам:

– Об'єм баз даних повинний бути практично не обмеженим (тобто обмежений лише параметрами апаратних засобів).

– БД повинна бути достатньо гнучкою для забезпечення проходження процесу перебудови по мірі її заповнення, так як попереднє проектування детальної структури бази даних є неможливим.

– БД повинні бути інтегровані з іншими БД та підтримувати не лише стандартні мультимедійні формати, але й ряд спеціальних гіпермедіа-середовищ (просторові структури молекул, хімічні структурні формули та ін.).

– Експлуатація та поповнення баз даних через комп'ютерні мережі має бути легко доступною та зрозумілою для користувачів, які не мають комп'ютерної підготовки (біологи, медики).

1.3. Характеристики біоінформатичних ресурсів

Однією з найвідоміших міжнародних організацій з тих, що створюють інструменти для аналізу інформації та здійснюють нагляд за наповненням баз даних біоінформатичного напряму, є Міжнародна система баз даних нуклеотидних послідовностей (International Nucleotide Sequence Database Collaboration (INSDC) – http://www.ncbi.nlm.nih.gov/collab), яка об'єднує три установи, нуклеотидні БД які мають різний набір сервісів:

БД EMBL-EBI у європейській лабораторії молекулярної біології (European Molecular Biology Laboratory, Geyzelberg, Germany) та Європейському інституті біоінформатики (European Bioinformatics Institute (EBI), UK); Заснована в 1984 році. https://www.ebi.ac.uk/. Європейська молекулярно-біологічна лабораторія (EMBL) – науково-дослідний інститут, який фінансується з коштів, що виділяються двадцятьма країнами-учасниками Європи і країною-партнером Австралією. Наукову діяльність у EMBL ведуть близько 85 незалежних груп, які охоплюють всі галузі молекулярної біології. Лабораторія складається з п'яти відділень: головна лабораторія в Гейдельберзі (Німеччина), філії в Греноблі (Франція), Гамбурзі (Німеччина), Монтерондо (передмістя Рима, Італія), а в Європейському інституті біоінформатики в Хінкстоні поблизу Кембріджа, Великобританія розташована сама БД EMBL-EBI.

БД GenBank National Centre for Biotechnology Information. GenBank знаходиться у Центрі біотехнологічної інформації США (National Center for Biotechnology Information (NCBI)), при національній медичній бібліотеці, яка є складовою частиною національного інституту здоров'я США. Прототип бази даних GenBank створено в 1979 році в Лос-Аламоській (Los–Alamos, штат Нью–Мексіко) та в 1988 році передана Національному інституту здоров'я. NCBI включає базу послідовностей ДНК (GenBank), базу даних статей в галузі

медицини та біології (PubMed), базу таксономічної інформації (TaxBrowser) та багато інших БД. Бази даних доступні через пошукову систему Entrez.

БД DDBJ – DNA Data Bank of Japan, Центру інформаційної біології (Center for Information Biology (CIB) при National Institute of Genetics, Japan), заснована в 1984 році. http://www.ddbj.nig.ac.jp/ .

БД GenBank, EMBL-EBI та DDBJ генетичних послідовностей, містять анотовані колекції всіх загально-доступних послідовностей ДНК, РНК та білків разом з літературними посиланнями. Поповнюються кожен день. Регулярно синхронізуються, тому можна вважати ці БД рівнозначними, відрізняються БД між собою інтерфейсом. Поповнюється ці БД безпосередньо авторами, що визначили первинну структуру фрагмента ДНК чи РНК. Кожна база даних має свій формат представлення інформації, крім того, деякі з них виникли досить давно та зараз рідко використовуються. Для переводу з формату в формат існують різні програми-конвертори. Крім форматів EMBL, GeneBank, Swiss-Prot та інших БД, орієнтованих на сприйняття інформації людиною, необхідно знати один з найбільш поширених комп'ютерно-орієнтованих форматів – FASTA-формат, який слугує для представлення нуклеотидних та амінокислотних послідовностей у вигляді, зручному для обробки на обчислювальних машинах. Важливо також вміти переводити в нього данні з більш високорівневих форматів.

Розглянемо більш докладно БД GenBank NCBI. Список ключових полів БД GenBank NCBI послідовностей (Nucleotide, Protein, EST, GSS) та опис відповідних ключових слів наведені в таблиці 1.2 та таблиці 1.3.

Таблиця 1.2 – Ключові поля для всіх секвенованих послідовностей бази даних (нуклеотидів, білків, EST, GSS).

Ключове слово	**Description**	**Опис**
Accession	The accession number assigned by NCBI.	Номер, присвоєний NCBI.
All Fields	All terms from all search fields in the database.	Усі терміни з усіх полів пошуку в базі даних.
Author	All authors from all references in the records.	Всі автори з усіх посилань у записах.
EC/RN Number	Enzyme Commission (EC) number for an enzyme activity.	Номер за класифікацією Комісії з ферментів – Enzyme Commission (EC) для активності ферментів.
Feature Key	Biological features listed in the Feature Table of the sequence records.	Біологічні особливості, перелічені в Таблиці характеристик записів послідовностей.
Filter	Filtered subsets of the database. An important kind of filter is based on the presence of links to other records.	Фільтровані підмножини бази даних. Важливий фільтр, заснований на наявності посилань на інші записи.
Gene Name	Gene names annotated on database records. For NCBI Reference Sequences, these names correspond to official nomenclature guidelines when possible. Submitters provide the gene names on GenBank/GenPept records.	Назви генів в записах баз даних. Для послідовностей NCBI ці назви відповідають офіційним рекомендаціям щодо номенклатури. Відправники надають назви генів у записах GenBank/ GenPept.
Genome Project	The numeric unique identifier for the genome project that produced the sequence records.	Числовий унікальний ідентифікатор для проекту геному, який створив записи послідовностей.

Продовження таблиці 1.2.

Ключове слово	Description	Опис
Issue	The issue number of the journals cited on sequence records, not generally useful in sequence databases.	Номер видання журналів, цитованих у записах послідовностей.
Journal	The name of the journals cited on sequence records. Journal names are indexed in the database in abbreviated form although many full titles are mapped to their abbreviations. Journals are also indexed by their by International Standard Serial Number (ISSN).	Назва журналів, цитованих у записах послідовностей. Назви журналів індексуються в базі даних у скороченому вигляді. Журнали також індексуються їх Міжнародним стандартним порядковим номером (ISSN).
Keyword	Keywords applied by submitter or from controlled vocabularies applied by NCBI or other databases. Except for specific kinds of records, such as the examples given below, the terms in this index are not well controlled. This field is unpopulated for many GenBank/GenPept records.	Ключові слова, застосовані суб'єктом або зі словників, застосованих NCBI або іншими базами даних. Це поле незаповнене для багатьох записів GenBank / GenPept.
Modification Date	The date of most recent modification of a sequence record. The date format is YYYY/MM/DD. Only the year is required. The Modification Date is often used as a range of dates. The colon (:) separates the beginning and end of a date range.	Дата останньої зміни запису послідовностей. Формат дати - РРРР / ММ / ДД. Дата модифікації часто зазначається як діапазон дат. Двокрапка (:) розділяє початок і кінець діапазону дат.

Продовження таблиці 1.2.

Ключове слово	**Description**	**Опис**
Molecular Weight	The molecular weight in Daltons of the protein chain calculated from the amino acids only. This may not correspond to the molecular weight of the protein obtained from biological samples because of incomplete data or post-translational modifications of the protein in living systems. The colon (:) separates the beginning and end of a molecular weight range.	Молекулярна маса білка в дальтонах розраховується з амінокислот. Може не відповідати молекулярній масі білка, отриманого з біологічних зразків через неповні дані або посттрансляційні модифікації білка в живих системах. Двокрапка (:) розділяє початок і кінець діапазону молекулярної маси.
Organism	The scientific and common names for the complete taxonomy of organisms that are the source of the sequence records. This vocabulary includes all available nodes in the NCBI taxonomy database.	Наукові та загальні назви для повної таксономії організмів, які є джерелом записів послідовностей. Цей словник включає всі доступні вузли в базі таксономії NCBI.
Page Number	The page numbers of the articles that are cited on the sequence record, not generally useful in sequence databases.	Номери сторінок статей, які цитуються у записі послідовностей.
Primary Accession	The primary accession number of the sequence record. This is the first one appearing on the ACCESSION line in the GenBank/GenPept format. Many records have additional secondary accessions representing records that have been merged. The Accession field indexes both primary and secondary accessions.	Первинний номер запису послідовностей. Це перше, що з’являється в рядку ACCESSION у форматі GenBank/GenPept. Багато записів мають додаткові вторинні приєднання, що представляють записи, які були об'єднані.
Primary Organism	The primary organism when there is more than one source organism.	Первинний організм, коли існує більше ніж одне джерело.

Продовження таблиці 1.2.

Ключове слово	**Description**	**Опис**
Properties	Molecular type, source database, and other properties of the sequence record. Terms indexed for this field are a useful classification system for sequence records.	Молекулярний тип, вихідна база даних та інші властивості запису послідовностей.
Protein Name	The names of protein products as annotated on sequence records.	Назви білкових продуктів, як зазначено в записах послідовностей.
Publication Date	The date that records were made public in Entrez. The date format is YYYY/MM/DD. The colon (:) separates the beginning and end of a date range.	Дата оприлюднення записів в Entrez. Формат дати - РРРР / ММ / ДД. Двокрапка (:) розділяє початок і кінець діапазону дат.
SeqID String	The NCBI identifier string for the sequence record. This is a brief structured format used by NCBI software.	Рядок ідентифікатора для запису послідовності. Це короткий структурований формат, який використо-вується програмним забезпеченням NCBI.
Sequence Length	The total length of the sequence − the number of nucleotides or amino acids in the sequence. The colon (:) separates the beginning and end of a length range.	Загальна довжина послідовності - кількість нуклеотидів або амінокис-лот у послідовності. Двокрапка (:) розділяє початок і кінець діапазону довжин.
Substance Name	The names of chemical substances associated with a record. This field is only populated for sequences extracted from structure records – PDB derived sequences.	Назви хімічних речовин, пов’язаних із записом. Це поле заповнюється лише для послідовностей, взятих зі структурних записів - послідовностей, отриманих PDB.
Text Word	Text on a sequence record that is not indexed in other fields.	Текст запису послідовно-стей, який не індексується в інших полях.

Кінець таблиці 1.2.

Ключове слово	**Description**	**Опис**
Title	Words and phrases found in the title of the sequence record. The title is the DEFINITION line of the GenBank/GenPept format of the record. This line summarizes the biology of the sequence and includes the organism, product name, gene symbol, molecule type, and sequence completeness.	Слова та фрази, знайдені в заголовку запису послідов-ностей. Назва - рядок DEFINITION формату GenBank/GenPept. Цей рядок підсумовує біологію послідовності та включає організм, назву продукту, символ гена, тип молекули та повноту послідовності.
Volume	Contains the volume number of the journals in references on the sequence record, not generally useful in the sequence databases.	Містить номери томів журналів у посиланнях на запис послідовностей.

Таблиця 1.3 – Ключові слова БД GenBank та опис відповідних (https://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html#ModificationsDateB).

Ключове слово	**Description**	**Опис**
Locus	The LOCUS field contains a number of different data elements, including locus name, sequence length, molecule type, GenBank division, and modification date.	Поле LOCUS містить ряд різних даних, включаючи назву локусу, довжину послідовності, тип молекули, розділ GenBank та дату модифікації.
Definition	Brief description of sequence; includes information such as source organism, gene name/protein name, or some description of the sequence's function.	Короткий опис послідовності; включає інформацію, таку як організм, назва гена/білка або опис функції послідовності.
Accession	The unique identifier for a sequence record.	Унікальний ідентифікатор для запису послідовностей.

Продовження таблиці 1.3.

Ключове слово	**Description**	**Опис**
Version	A nucleotide sequence identification number that represents a single, specific sequence in the GenBank database.	Ідентифікаційний номер нуклеотидної послідовності, який представляє єдину конкретну послідовність у базі даних GenBank.
Keywords	Word or phrase describing the sequence. If no keywords are included in the entry, the field contains only a period.	Слово або фраза, що описують послідовність. Якщо до запису не включені ключові слова, поле містить лише період.
Source	Free-format information including an abbreviated form of the organism name, sometimes followed by a molecule type.	Інформація вільного формату, включає скорочену назву організму, іноді супроводжується типом молекули.
Organism	The formal scientific name for the source organism (genus and species, where appropriate) and its lineage, based on the phylogenetic classification scheme used in the NCBI Taxonomy Database.	Офіційна наукова назва організму (рід та вид, де це доцільно) та його походження, заснована на філогенетичній класифікації, що використовується у базі даних таксономії NCBI.
Reference	Publications by the authors of the sequence that discuss the data reported in the record. References are automatically sorted within the record based on date of publication, showing the oldest references first.	Публікації авторів, в яких обговорюються дані, зазначені у записі. Посилання автоматично сортуються в межах запису за датою публікації, спочатку показуючи найдавніші посилання.
Authors	List of authors in the order in which they appear in the cited article.	Список авторів у тому порядку, в якому вони з’являються у цитованій статті.
Title	Title of the published work or tentative title of an unpublished work.	Назва опублікованого твору або попередня назва неопублікованого твору.
Journal	MEDLINE abbreviation of the journal name.	Абревіатура від назви журналу.
PubMed	PubMed Identifier (PMID).	PubMed Ідентифікатор (PMID).

Кінець таблиці 1.2.

Ключове слово	**Description**	**Опис**
Features	Information about genes and gene products, as well as regions of biological significance reported in the sequence. These can include regions of the sequence that code for proteins and RNA molecules, as well as a number of other features.	Інформація про гени та продукти генів, а також про ділянки, які мають біологічне значення. Сюди можна віднести області послідовності, які кодують білки та молекули РНК, а також ряд інших ознак.
Origin	The ORIGIN may be left blank, may appear as "Unreported," or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available).	ORIGIN може бути порожнім, може відображатись як «Неповідомлений», або може давати локальний покажчик на початок послідовності, як правило, включаючи експериментально визначений сайт рестрикції або генетичний локус (якщо він є).
//	Termination line.	Мітка кінця запису.

Розглянемо більш докладно формат представлення інформації в БД GenBank. Картка GenBank, представляє з себе текстовий файл з обмеженою довжиною строки. Ключові слова, що визначають характер представленої в різних областях картки інформації є жорстко заданими. Для картки GenBank ключові слова створюють двосимвольні послідовності на початку строки та служать для розширення можливостей БД. Список ключових слів наведено в таблиці 1.3.

>gi|15865658|ref|AAL09996.1| magnetosome protein MamA [*Magnetospirillum gryphiswaldense* MSR-1]

MSSKPSNMLDEVTLYTHYGLSVAKKLGANMVDAFRSAFSVNDDIRQVYY
RDKGISHAKAGRYSEAVVMLEQVYDADAFDVEVALHLGIAYVKTGAVDR
GTELLERSIADAPDNIKVATVLGLTYVQVQKYDLAVPLLVKVAEANPVNF
NVRFRLGVALDNLGRFDEAIDSFKIALGLRPNEGKVHRAIAYSYEQMGSHE
EALPHFKKANELDERSAV

Назвою цієї послідовності є:

" gi | 15865658 | ref | AAL09996.1 |".

Послідовності записуються у вигляді нуклеїнових кислот або амінокислот, в них допускаються пропуски і символи вирівнювання. Складові елементи кодуються загальноприйнятими однолітерними кодами (IUB / IUPAC), при цьому додатково дозволено використовувати символи нижнього регістра, дефіс для пропусків, і символи «U» і «*» в амінокислотних послідовностях. Числа не допускаються, але використовуються в деяких базах даних для позначення позиції.

Для запису амінокислот існує 24 звичайних кода та 3 спеціальних (табл. 1.4).

Таблиця 1.4 – Загальноприйнятий однолітерний код для запису амінокислот.

Код амінокислоти	**Значення**
A	Аланін
B	Аспарагінова кислота (D) або Аспарагін (N)
C	Цистеїн
D	Аспарагінова кислота
E	Глутамінова кислота
F	Фенілаланін
G	Гліцин
H	Гістидин
I	Ізолейцин

Кінець таблиці 1.4

Код амінокислоти	**Значення**
J	Лейцин (L) або Ізолейцин (I)
K	Лізин
L	Лейцин
M	Метіонін
N	Аспарагін
O	Піролізин
P	Пролін
Q	Глутамін
R	Аргінін
S	Серин
T	Треонін
U	Селеноцистеїн
V	Валін
W	Триптофан
Y	Тирозин
Z	Глутамінова кислота (E) або Глутамін (Q)
X	будь-який
*	зупинка трансляції
-	пропуск невизначеної довжини

Для запису нуклеотидних послідовностей ДНК за рекомендацією IUPAC містить символи латинського алфавіту.

A = аденін

C = цитозин

G = гуанін

T = тимін

R = G A гуанін або аденін – пуринові (purine) основи

Y = T C перемединові (pyrimidine) основи

K = G T містять кето групу

M = A C містять аміногрупу

S = G C «сильні» (strong), утворюють 3 водневі зв'язки

W = A T «слабкі» (weak), утворюють 2 водневі зв'язки

B = GTC (будь-який, крім A)

D = GAT (будь-який, крім C)

H = ACT (будь-який, крім G)

V = GCA (будь-який, крім T)

N = AGCT (будь-який)

Символи, що позначають невизначеність послідовності, необхідні, коли точна послідовність невідома, несуттєва, або відомі різні її варіанти. Для запису послідовностей РНК зазвичай досить символів A, C, G, U (уридин). Всі послідовності записуються без пробілів, зліва направо, від 5᾽- кінця до 3᾽ - кінця.

В 1988р. стартував проект «Геном людини». Вже при розробці проекту було прийнято критично важливі рішення, що суттєво вплинуло на подальший розвиток геноміки та біоінформатики. Перше з них полягало у тому, щоб секвенувати не тільки геном людини, а і геноми модельних організмів: нематоди *Caenorhabditis elegans*, плодової мухи *Drosophila melanogaster*, дріжджів *Saccharomyces cerevisiae*, рослин *Arabidopsis thaliana,* бактерій *Escherichia coli. Bacillus subtilis* та інших. При виборі об’єктів секвенування в основному враховувався баланс між вивченістю організму та розміром його генома. В результаті став можливий порівняльний аналіз одразу багатьох геномних даних. Друге таке ж важливе рішення полягало у тому, що данні секвенування геномів одразу ж ставали доступними світовому науковому співтовариству. В 1996 р. було сформульовано «Бермудські принципи» (названі за місцем проведення конференції), згідно з якими навіть невеликі фрагменти геномів, отримані в рамках проекту «Геном людини» і аналогічних програм, одразу ж розміщались у банках даних і могли бути використані усіма бажаючими. Одночасно в журналах публікувалися результати аналізу великих

секвенованих фрагментів геномів і цілих хромосом. При цьому, більшість провідних журналів ввели у практику неприйняття до публікації матеріалів, не розміщених хоча б в одну з загальнодоступних БД.

Запитання до розділу 1

1. Дати визначення бази даних (БД).
2. Який перший важливий з біологічної точки зору результатом, було отримано за допомогою аналізу послідовностей?
3. Які функції виконують БД ?
4. З якими об’єктами оперують БД?
5. Яким чином, зазвичай, класифікують БД?
6. На які типи за тематикою поділяються БД?
7. Наведіть формати представлення інформації в БД.
8. Які Ви знаєте основні вимоги до програмного забезпечення баз даних?
9. Наведіть приклади архівних БД та БД, що куруються.
10. Які методики пошуку інформації у БД Вам відомі?
11. Розкрийте суть роботи програми BLAST.
12. Перерахуйте найбільш відомі біоінформатичні ресурси.
13. Основні характеристики БД нуклеотидних послідовностей EMBL.
14. Що Вам відомо про БД білкових послідовностей та 3d структур?
15. Які біоінформатичні ресурси можна віднести до спеціалізованих?
16. Яким основним вимогам повинно задовольняти програмне забезпечення баз даних.

Література до розділу 1

1. Bioinformatics: Sequence, structure and database – Oxford University Press, 2001.
2. URL: http://anil.cchmc.org/University.html
3. Lesk M. Introduction to Bioinformatics – Oxford University Press Inc. New York, 2002.
4. Waterfield M. Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. //Nature. 1983, 304, 35-39.
5. Арчаков А.И. Геномика, протеомика и биоинформатика – науки XXI столетия //Фармацевтический вестник №9 (208), 2001.
6. Глазко В.И., Глазко Г.В. Введение в генетику, биоинформатика, ДНК-технология, генная терапия, ДНК-экология, протеомика, метаболика – К.:КВІЦ, 2003.
7. Дурбин Р., Эдди Ш., Крог А., Митчисон Г. Анализ биологических последовательностей. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2006. – 480 с.
8. Игнасимуту С., Основы биоинформатики, [пер. с англ. А.А. Чумичкина], М.-Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 320 с.
9. Сетубал Ж., Мейданис Ж. Введение в вычислительную и молекулярную биологию. – М. – Ижевск: НИЦ «Регулярная и хаотическая динамика», Институт компьютерных исследований, 2007. – 420 с.

Розділ 2. БД БІЛКОВИХ ПОСЛІДОВНОСТЕЙ

У 1971 р. заснований Банк даних білків Едгаром Меєром (Edgar Meyer) і Валтером Гамільтоном (Walter Hamilton), співробітниками Брукгевенської національнрї лабораторії (Brookhaven Protein Data Bank в Brookhaven National Laboratory (BNL) – одна з 16 національних лабораторій Міністерства енергетики США, дослідження в області ядерної фізики та молекулярної біології.

Міжнародна організація Всесвітній Банк даних білків (PDB) складається з організацій по всьому світу, що займаються внесенням даних до бази даних PDB та розповсюдженням накопиченої інформації. Членами організації PDB зараз є RCSB PDB (Research Collaboratory for Structural Bioinformatics, співробітництво у дослідженні структурної біоінформатики, http://www.rcsb.org/, США), PDBE (Protein Data Base Europe, Європа) і PDBJ (Protein Data Base Japan, Японія). Місією PDB є підтримка єдиного архіву даних всіх структур біологічних макромолекул та вільне розповсюдження цієї інформації. Крім того, організацією підтримуються та приводяться до спільного формату багато інших баз даних, що містять інформацію щодо функції білків та їх еволюції.

Крім того, до найбільш відомих БД білків відносяться:

UniProt – це консорціум, який складається з наукових колективів Європейського інституту біоінформатики (EBI), Швейцарського інституту біоінформатики (SIB) та білкового інформаційного ресурсу (PSD-PIR). Консорціум UniProt є центральним ресурсом білкових послідовностей та функціональних анотацій для біоінформатичних досліджень білків. Бази даних консорціуму UniProt містять інформацію про послідовності протеїнів та дані щодо їх функцій, отримані з проектів сиквенсу геномів. БД містять також великий масив даних про біологічні функції протеїнів з журнальних статей, що містять оригінальні дослідження.

БД **Swissprot** Швейцарського інституту біоінформатики (Swiss Institute of Bioinformatics (SIB), http://www.isb-sib.ch/ містить анотовані амінокислотні послідовності, трансльовані з нуклеотидних послідовностей EMBL; адаптовані послідовності з PSD-PIR; а також послідовності опубліковані в літературі і надіслані безпосередньо авторами. БД **Swissprot** містить високоякісні анотації без збиткової інформації, посилання на споріднені бази даних (EMBL, GenBank, PROSITE**,** PDB). Кожна анотація містить опис функції білка, його доменної структури, особливостей пост-трансляційної модифікації. Оновлюється щотижня. Для академічних користувачів є безкоштовною.

БД PSD-PIR (Protein Sequence Database-Protein Information Resource, National Biomedical Research Foundation, Національний фонд медико-біологічних досліджень, США, http://www-nbrf.georgetown.edu/pirwww/dbinfo/irpsd.html) містить інформацію щодо білків, для яких відомі нуклеотидні послідовності. Пошук організовано як по таксономії так і гомології. Має низький рівень зайвої інформації. Поповнюється щотижня.

БД MMDB (Molecular Modelling Database, США, http://www.ncbi.nlm.nih.gov/Structure/) містить просторові структури білків, визначені дослідним шляхом (рентгеноструктурною кристалографією та ЯМР-спектроскопією), надає інформацію про біологічну функцію та механізми, що з нею пов’язані; еволюційну історію і взаємозв’язок між макромолекулами. Входить до складу PDB, що містить також теоретичні моделі. Всі структури цієї бази даних мають первинні структури в NCBI. Оновлюється щоденно.

ENZYME (http://www.expasy.ch/enzyme**) –** БД, що містить інформацію щодо номенклатури ферментів, і описує всі типи білків, яким присвоєно номер ЕС (Enzyme Commission). Пошук реалізовано по ЕС-номеру, класам ферментів, хімічним компонентам, по кофакторам, по назвам хвороб, пов’язаних з ферментом.

2.1 База даних UniProt

База даних UniProt (UniProtKB) є сховищем даних для функціональної інформації про послідовності білків з докладною та точною анотацією (назва або опис білка, таксономічна інформація, класифікація, перехресне посилання та цитування літератури). База даних UniProt, зокрема UniProtKB/Swiss-Prot, використовується для доступу до функціональної інформації про білки. Кожен запис UniProtKB містить послідовність амінокислот, назву або опис білка, таксономічні дані та інформацію про цитування. Сюди входять загально прийняті біологічні онтології, класифікації та перехресні посилання, а також чіткі вказівки на якість анотації.

UniProtKB складається з двох частин: UniProtKB/Swiss-Prot та UniProtKB/TrEMBL. Перша містить високоякісні анотації білків, зроблені вручну. Анотація робиться вручну, складається з аналізу, порівняння та об'єднання всіх доступних послідовностей для даного білка, а також критичний огляд супутніх експериментальних та прогнозованих даних. UniProtKB/Swiss-Prot прагне надати всю відому релевантну інформацію про конкретний білок. В одному записі описано різні білкові продукти, отримані з певного гена даного виду, білкові сімейства та групи регулярно переглядаються.

UniProtKB/TrEMBL містить високоякісний комп'ютерний аналіз записів, збагачений автоматичною анотацією та класифікацією.

Архів UniProt БД UniParc (UniProt Archive; архів бази даних UniProt) є базою даних архівних білків з усіх основних загальнодоступних ресурсів. UniParc – це найповніша загальнодоступна база даних білкових послідовностей, що забезпечує посилання на всі основні джерела та версії цих послідовностей. UniParc містить білкові послідовності та перехресні посилання на інші бази даних. UniParc призначений для збору всіх загальнодоступних даних про послідовності білків і містить усі послідовності білків з основних

загальнодоступних баз даних білкових послідовностей. Таким чином, **UniParc є найповнішою загальнодоступною базою даних, що не має зайвих білкових послідовностей**. Послідовність білка може існувати в декількох базах даних і не один раз повторюватися у окремій базі даних, таким чином створюючи зайву інформацію. UniParc долає цю проблему, зберігаючи кожну унікальну послідовність лише один раз та призначаючи їй унікальний ідентифікатор UniParc. UniParc обробляє всі послідовності просто як текстові рядки – послідовності, які на 100% однакові по всій довжині, об'єднуються незалежно від того, є вони одного чи іншого виду.

Три бази даних UniRef – UniRef100, UniRef90 і UniRef50 – автоматично об'єднують послідовності. UniRef100 базується на всіх записах UniProtKB. UniRef100 об'єднує ці записи за повною ідентичністю послідовностей. Ідентичні послідовності та субфрагменти послідовностей представлені у вигляді єдиного запису в UniRef100. UniRef90 і UniRef50 представляють записи із взаємною ідентичністю послідовностей на 90% або більше та, відповідно, на 50% або більше, з посиланнями на відповідні записи в UniProtKB. Бази даних UniRef забезпечують кластеризовані набори послідовностей з UniProtKB та вибраних записів UniParc, щоб забезпечити повне охоплення послідовностей. UniRef90 та UniRef50 дозволяють зменшити розмір бази даних приблизно на 40% та 65% відповідно, забезпечуючи значно швидший пошук послідовностей.

2.1.1 Зростання кількості послідовностей в UniProt

Портал протеомів UniProtKB забезпечує доступ до протеомів понад 84 тис. видів з повністю секвенованими геномами. Більшість цих протеомів засновані на трансляції послідовностей геномів у бази даних INSDC (International Nucleotide Sequence Database Collaboration) – European Nucleotide Archive (ENA), GenBank та DDBJ. При останньому оновленні у 2018 році до бази даних було включено

набір протеомів приматів, десяток геномів морських ссавців за допомогою бази даних RefSeq.

Постійне зростання кількості геномів, що секвенуються, є проблемою для баз даних, оскільки значна частина цього зростання характеризується секвенуванням дуже схожих і майже однакових послідовностей (90% білків одного виду мають > 90% ідентичність). Розробники бази даних слідкують за зростанням кількості секвенованих геномів, що надає їм змогу керувати кількістю інформації. Процес видалення надлишкових послідовностей був вперше введений у 2015 році. Цей процес виявляє та видаляє майже однакові протеоми одного виду до того як включити їх в UniProtKB та розміщує їх послідовності в UniParc. В даний час у результаті цього процесу з UniProtKB вилучено 38% всіх протеомів (241 мільйон білків). Як видно з рис.2.1, зниження кількості надмірних послідовностей значно зменшило розмір UniProtKB, а також зробило її зростання більш масштабованим. Цей підхід зараз поширився від прокаріотів до грибів, що призвело до знецінення ~1 мільйона записів білків грибів у 2016 році.

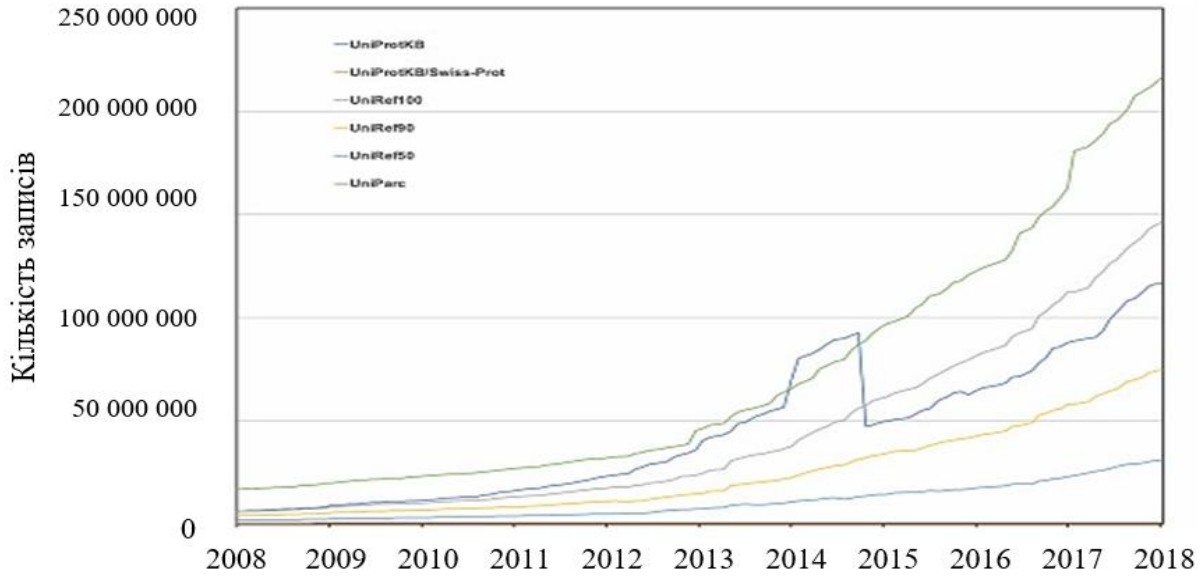


Рисунок 2.1. Кількість записів у базі даних за останні 10 років

2.1.2 Інформація в UniProt, що забезпечується кураторами

Завантаження даних із літератури є критично важливим для бази даних UniProt. Інформація з наукових публікацій, зберігається у UniProtKB/Swiss-Prot та описує функціональну інформацію як у формі зручних текстових даних, так і за допомогою структурованих словників, таких як генетична онтологія (GO) або ChEBI (Chemical Entities of Biological Interest, словник молекулярних об'єктів, орієнтований на невеликі хімічні сполуки). Цей процес є трудомістким, куратори збирають та оцінюють велику кількість даних з відповідних публікацій, але ця процедура є найбільш ефективним методом вилучення всіх релевантних даних з паперових носіїв.

Записи UniProtKB/Swiss-Prot служать джерелом функціональних даних для розробки та вдосконалення інструментів прогнозування біоінформатики, тому автори надають пріоритетну роль у наданні функціональних даних, які в даний час неможливо передбачити обчислювальними методами. Наприклад, білки, які б за передбаченнями можна було б віднести до ферментів, але насправді вони є нефункціональними через втрату специфічних амінокислотних залишків. Тому важливо переконатися, що у UniProtKB/Swiss-Prot наявна достатньо широка колекція анотованих білків, щоб надати цінність новим записам, оскільки таксономічний діапазон повністю секвенованих протеомів продовжує розширюватися.

Після того, як запис переміщено в UniProtKB/SwissProt з UniProtKB/TrEMBL, необхідно запис регулярно оновлювати, щоб забезпечити поточну інформацію про білок, яка з'являється в науковій літературі. Це складне завдання, оскільки нові знання про функції та більш уточннені експериментальні результати з'являються постійно і при цьому можуть мати суперечливі дані.

2.1.3 Автоматична анотація в UniProt

Автоматична анотація геномів у UniProt має дві допоміжні системи прогнозування на основі правил UniRules та системи автоматичної статистичної анотації (SAAS). Ці системи передбачення можуть здійснювати анотації властивостей білка, такі як назви білків, функції, каталітичну активність, приналежність до біосинтетичного шляху та розташування в клітині, поряд із специфічною для послідовності інформацією, такою як позиції посттрансляційних модифікацій та активні сайти. Кількість правил, що використовуються для анотації, складає у 2018 році понад 6000, як показано на рисунку 2.2.

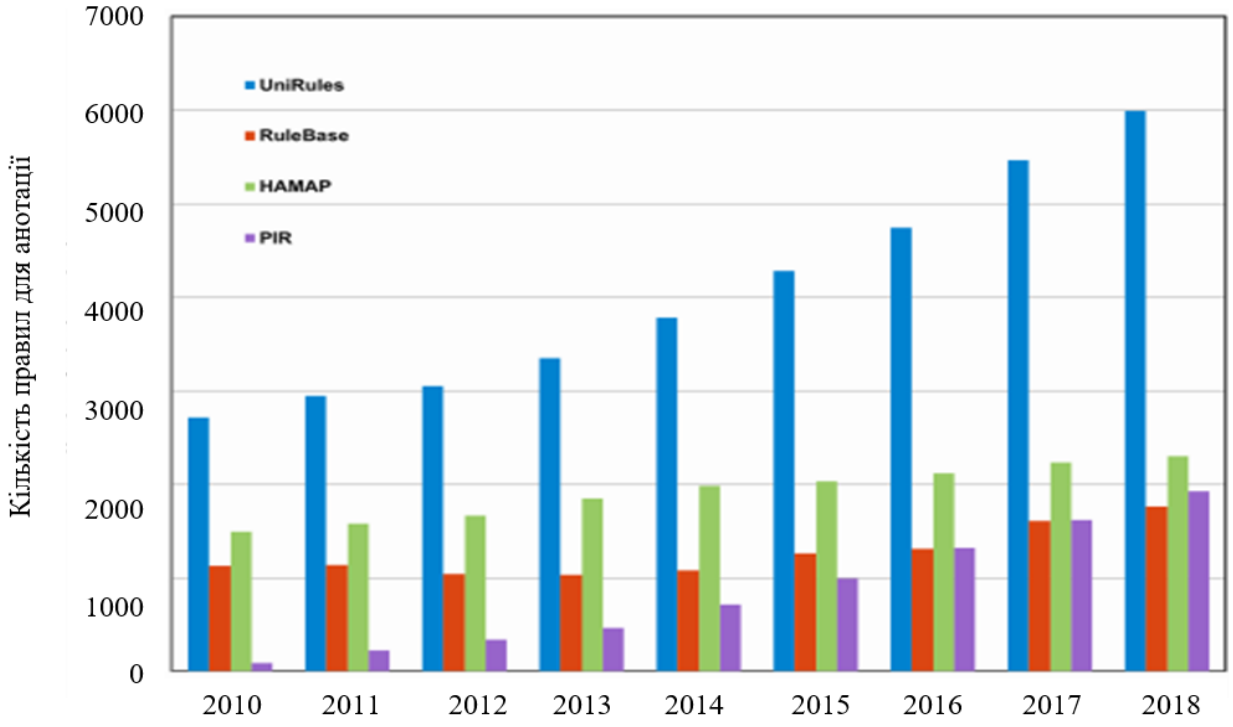


Рисунок 2.2 – Кількість правил для автоматичної анотації у базах даних UniProt.

2.1.4 Пан-протеоми в UniProt

Щоб доповнити зростаючий набір протеомів, UniProt розробив пан-протеоми, аналогічні за концепцією до пан-геномів. Пан-протеом – це повний набір білків, експресований групою сильно споріднених організмів (наприклад, декілька штамів одного і того ж виду бактерій). Пан-протеоми забезпечують репрезентативний набір всіх послідовностей в межах таксономічної групи і фіксують унікальні послідовності, не знайдені в протеомі даної групи. Пан-протеоми UniProtKB охоплюють усі не надлишкові протеоми і спрямовані на користувачів, зацікавлених у філогенетичних порівняннях та вивченні еволюції геномів та різноманітності генів.

Для кожного референтного кластеру протеому, також відомого як репрезентативна група протеомів, пан-протеом – це сукупність усіх послідовностей у референтному протеомі плюс унікальні білкові послідовності, які зустрічаються в інших видів або штамів кластера, але не в референтному протеомі. Пан-протеоми доступні у вигляді файлів послідовностей форматів FASTA на веб-сайті FTP. На веб-сайті UniProt є посилання на завантаження файлу пан-протеому.

2.1.5 Оновлення на веб-сайті UniProt

За останній (2018) рік було додано три нові візуалізації на веб-сайт UniProt. По-перше, це метод для перегляду молекулярних взаємодій, по-друге, спосіб перегляду субклітинної локалізації білків і, нарешті, перегляд молекулярної структури. Вони разом дозволяють користувачам швидко розуміти молекулярний контекст записів UniProt. Для записів UniProtKB, які містять розділ Interaction (Взаємодія), показані деталі бінарних взаємодій білка з іншими білками, використовуючи високоякісний набір даних, наданий консорціумом

IMEx (International Molecular Exchange Consortium). Бінарні взаємодії білка показані як матриця, яка показує партнерів взаємодії досліджуваного білка, а також показує, хто з цих партнерів взаємодіє один з одним.

Одним із типів інформації про білок є його субклітинне розташування. Цей розділ надає інформацію про розташування та топологію зрілого білка в клітині. Користувачі мають змогу візуально досліджувати клітинне розташування білка. Візуалізація представляє шаблони зображень із ресурсу COMPARTMENTS у поєднанні з даними про розташування білка від UniProt (експертна анотація, автоматична анотація на основі правил) та імпортованих анотацій із ресурсу GO.

Структурна інформація важлива для розуміння молекулярних механізмів, які дозволяють білкам виконувати свої специфічні функції. UniProt забезпечує перегляд білкової структури в розділі Structure (Структура), а також у програмі перегляду білка ProtVista (рис. 2.3). Структури відображаються за допомогою програми перегляду Litemol.

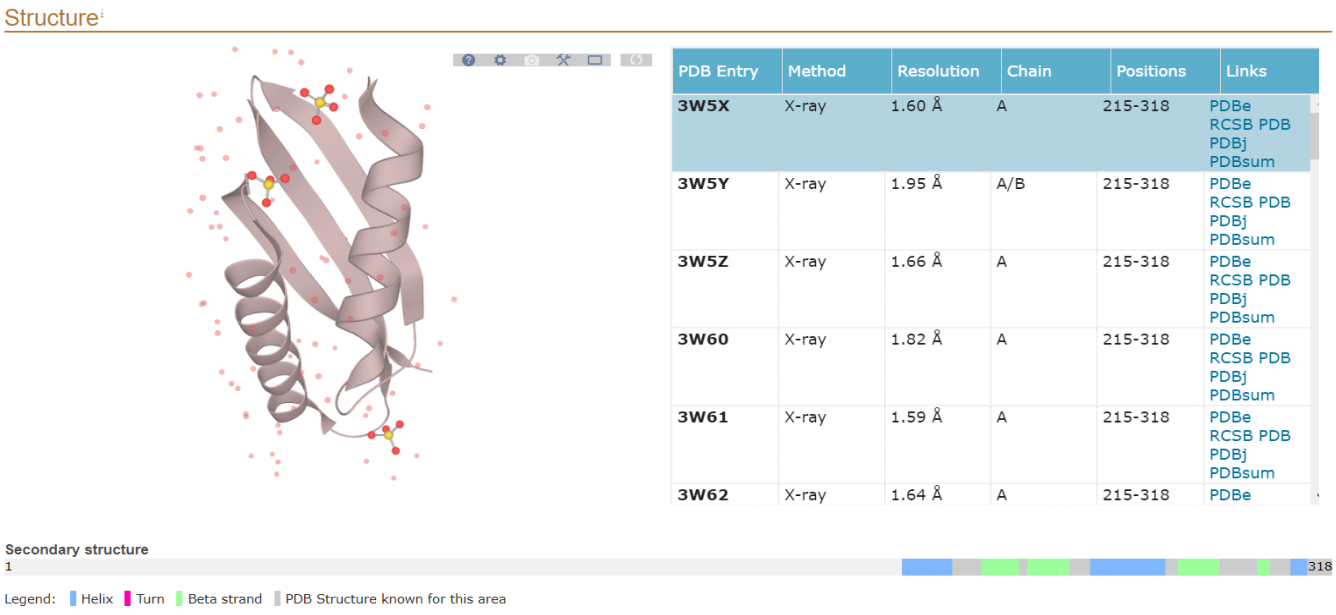


Рисунок 2.3 – Молекулярна структура білка MamM *Magnetospirillum gryphiswaldense* MSR-1 (Primary accession number: V6F235), показано у програмі перегляду білка ProtVista.

2.1.6. Основні можливості бази даних UniProt

База даних UniProt представляє більшу частину інформації про даний білок, зокрема його функції, структуру, сайти на поверхні, посттрансляційні модифікації, зв'язок із захворюваннями тощо. Зображення стартової сторінки наведено на рисунку. Найбільша увага надається тим модулям консорціуму UniProt, описи яких подано у спеціальних рамках, що виділені різними кольорами (рис. 2.4).

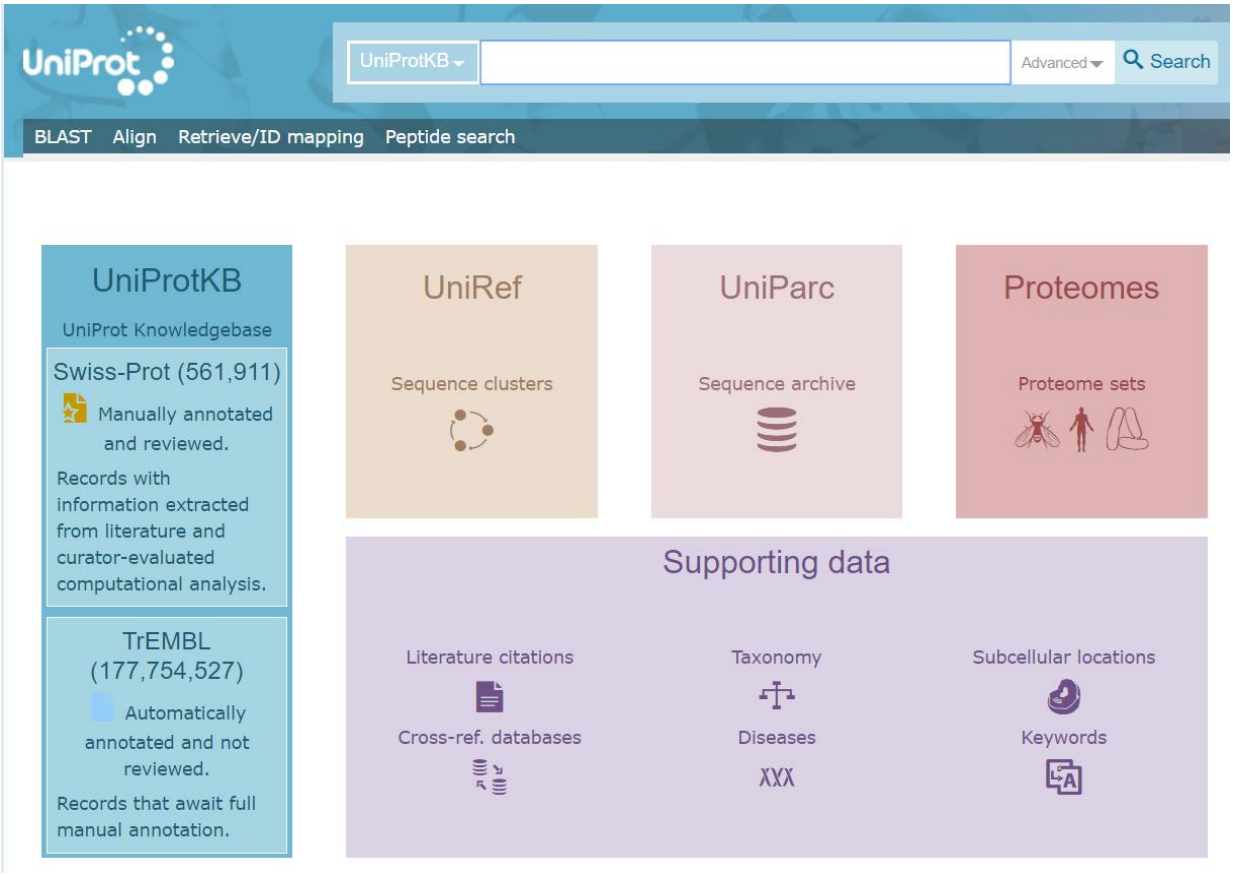


Рисунок 2.4 – Стартова сторінка бази даних UniProt

Зверху сторінки знаходиться рядок пошуку, за допомогою якого можна знайти потрібний білок. Однак, це можна зробити лише за допомогою

спеціального шифру, що позначає даний білок саме у цій базі даних. Так, наприклад, для білка магнітосомної мембрани *Magnetospirillum gryphiswaldense* MSR-1 цей шифр – MAMM_MAGGM.

При виборі необхідного білка відкривається сторінка, на якій містяться основні дані про нього. Зверху сторінки знаходяться дані про назву білка, ген, що його кодує, організм, а також статус анотованості амінокислотної послідовності білка. Після цієї інформації знаходяться дані про функції білка. Важливим є те, що у лівій частині сторінки можливо вибрати ті чи інші дані, що можна додати, чи, навпаки, виключити із сторінки. Це може бути корисним для дослідників, які шукають лише певну інформацію про білок (взаємодії з іншими білками, пов'язані хвороби тощо), це значно економить час і дозволяє отримати лише необхідні дані (рис. 2.5).

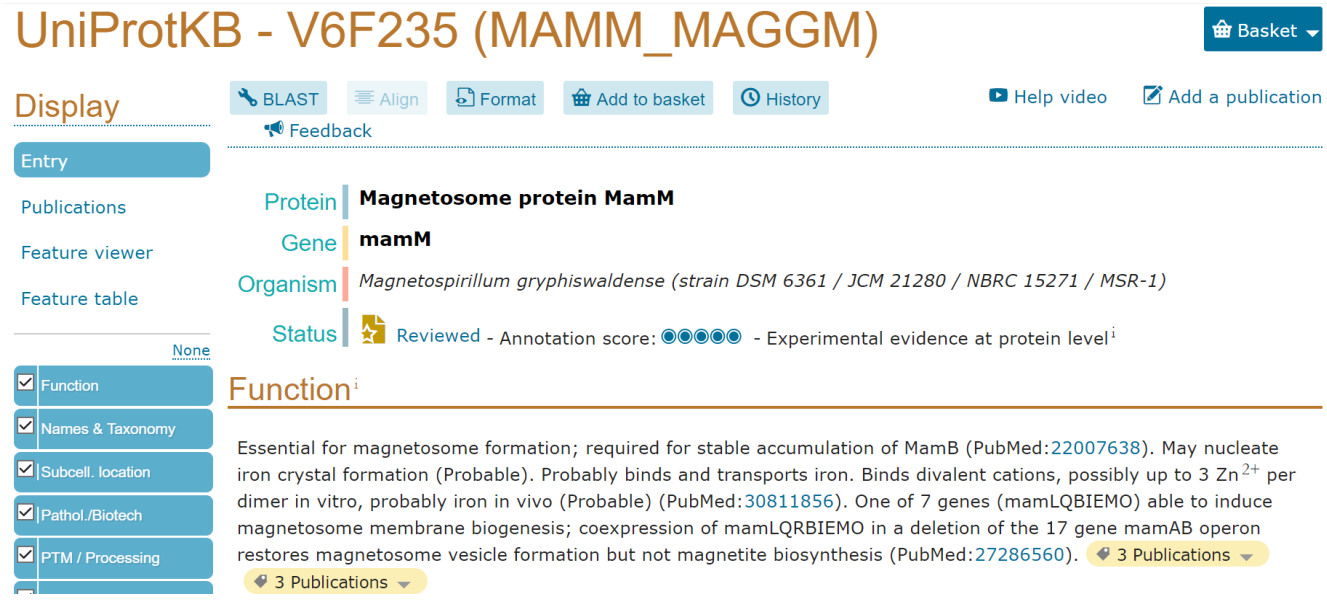
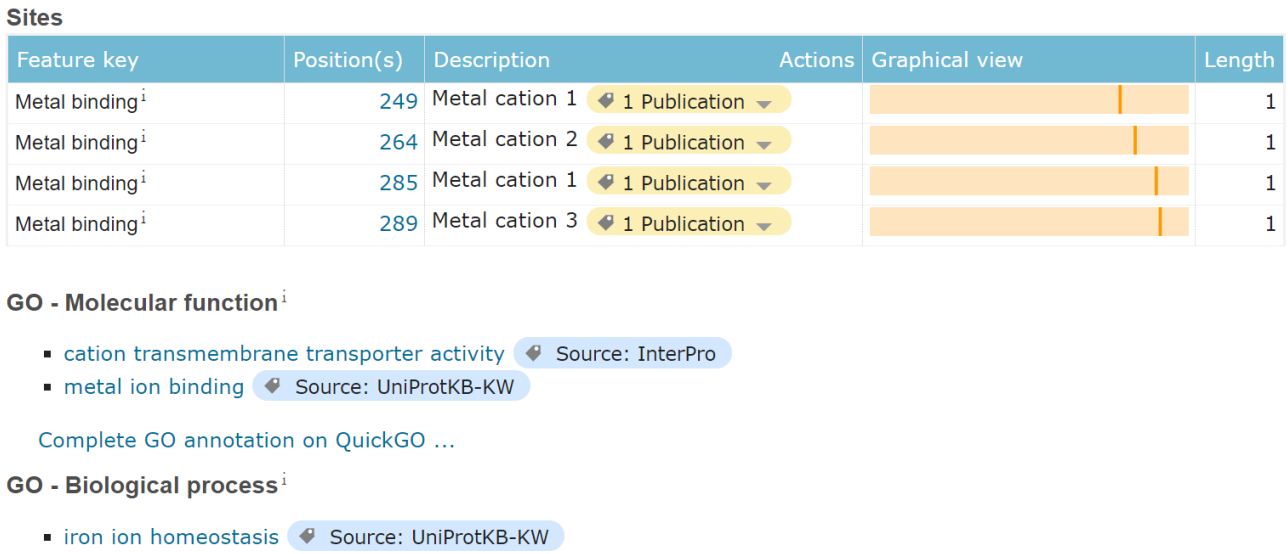


Рисунок 2.5 – Основна інформація про білок: первинні дані та функції

Після цієї інформації містяться дані про сайти на поверхні білка. Для обраного білка магнітосомного острівця це, в основному, металзв'язуючі сайти.

Також наведені дані про функції та процеси, в яких даний білок бере участь. Кожна з перелічених функцій містить посилання на першоджерела – інші бази даних (рис. 2.6).



Sites

Feature key	Position(s)	Description Actions	Graphical view	Length
Metal binding	249	Metal cation 1 — 1 Publication		1
Metal binding	264	Metal cation 2 — 1 Publication		1
Metal binding	285	Metal cation 1 — 1 Publication		1
Metal binding	289	Metal cation 3 — 1 Publication		1

GO - Molecular function

- cation transmembrane transporter activity — Source: InterPro
- metal ion binding — Source: UniProtKB-KW

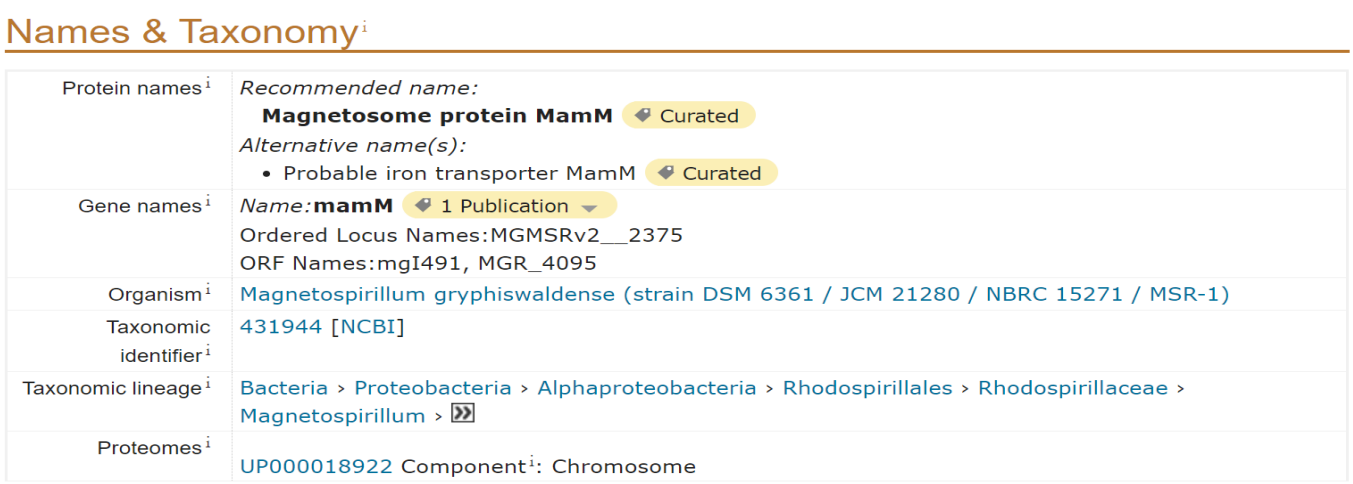
Complete GO annotation on QuickGO ...

GO - Biological process

- iron ion homeostasis — Source: UniProtKB-KW

Рисунок 2.6 – Основна інформація про білок: сайти на поверхні, молекулярна функція білка та процеси, в яких він бере участь.

Також у даній базі даних міститься інформація про назви білка, взята з інших баз даних чи літератури. Крім того, подано дані про таксономічне положення поліпепитиду (рис. 2.7).



Names & Taxonomy

Protein names	*Recommended name:* **Magnetosome protein MamM** — Curated *Alternative name(s):* • Probable iron transporter MamM — Curated
Gene names	*Name:***mamM** — 1 Publication Ordered Locus Names:MGMSRv2__2375 ORF Names:mgI491, MGR_4095
Organism	Magnetospirillum gryphiswaldense (strain DSM 6361 / JCM 21280 / NBRC 15271 / MSR-1)
Taxonomic identifier	431944 [NCBI]
Taxonomic lineage	Bacteria › Proteobacteria › Alphaproteobacteria › Rhodospirillales › Rhodospirillaceae › Magnetospirillum › »
Proteomes	UP000018922 Component: Chromosome

Рисунок 2.7 – Основна інформація про білок: назви та таксономія

У базі даних UniProt міститься інформація про розташування білка у клітині чи міжклітинному просторі (рис. 2.8).

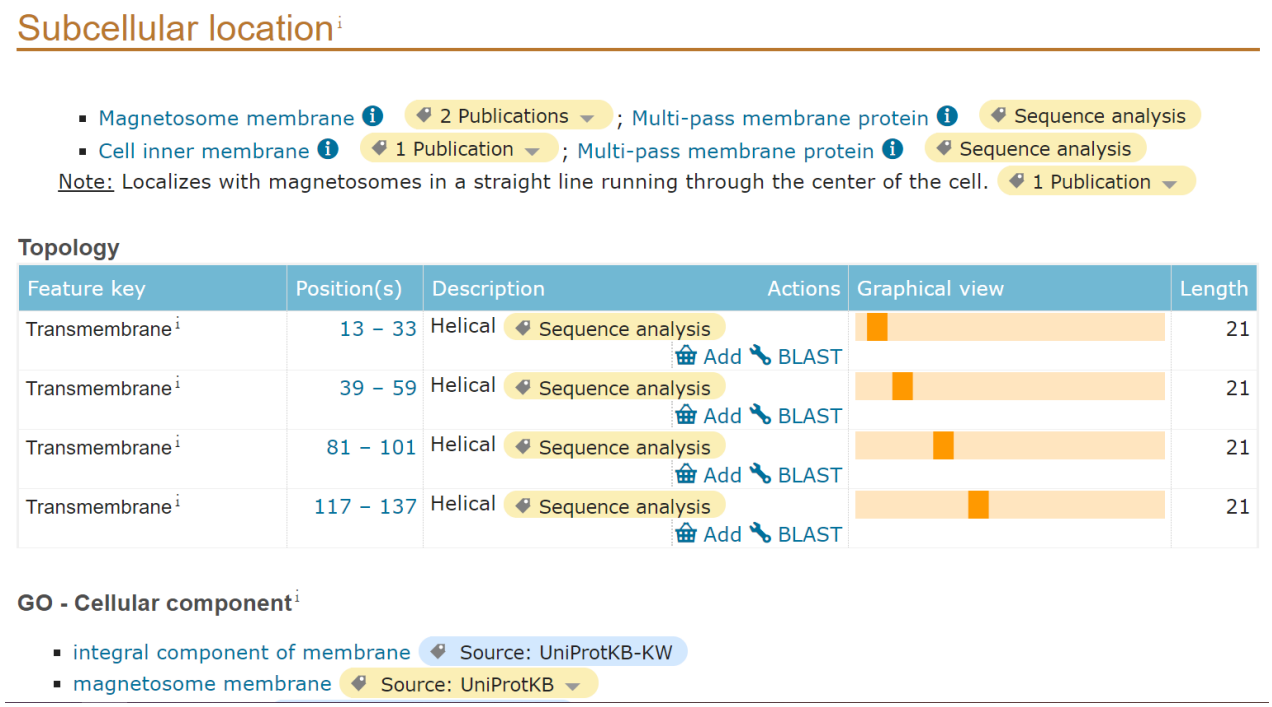


Рисунок 2.8 – Основна інформація про білок: розташування у клітині чи позаклітинному просторі

Наступною інформацією про білок є дані про те, які хвороби можуть бути спричинені внаслідок неправильного функціонування даного білка. Найчастіше такі помилки виникають через мутації, тому нижче міститься таблиця із можливими варіантами мутагенезу у досліджуваній поліпептидній послідовності (рис. 2.9).

Pathology & Biotech

Disruption phenotype

Single gene disruption has no growth defects, no accumulation of magnetite, forms empty intracellular magnetosome vesicles, decreased levels of MamB, mislocation of MamC in 1-3 foci or rarely in a shortened chain (PubMed:22007638). Magnetosome vesicles are fewer and smaller, aligned in a chain with the filament, only a few have very small crystals. Other, possibly precursor magnetosome vesicles are visible. MamI mislocalized to cell inner membrane or in 1 to a few patches (PubMed:27286560). Deletion of approximately 80 kb of DNA, including this operon, leads to cells that are non-magnetic, lack internal membrane systems, grow poorly, have reduced mobility and take-up and accumulate iron poorly (PubMed:13129949). 3 Publications

Mutagenesis

Feature key	Position(s)	Description Actions	Graphical view	Length
Mutagenesis	6 – 9	CAVC → SAVS: Wild-type magnetic response. 1 Publication		4
Mutagenesis	9	C → S: Wild-type magnetic response. 1 Publication		1
Mutagenesis	46	Y → D: Loss of magnetic response. 1 Publication		1
Mutagenesis	46	Y → H: About 90% magnetic response. 1 Publication		1

Рисунок 2.9 – Основна інформація про білок: хвороби та патології

Досить важливою інформацією про білок є дані про його процесинг та посттрансляційні модифікації. Ці дані також містяться у базах даних UniProt, разом із посиланнями на джерела досліджень, або на схожий білок, за яким ці модифікації були передбачені (рис. 2.10).

PTM / Processing

Molecule processing

Feature key	Position(s)	Description Actions	Graphical view	Length
Chain (PRO_0000447739)	1 – 318	Magnetosome protein MamM Add BLAST		318

Expression

Induction

Part of the probable 17 gene mamAB operon. 1 Publication

Interaction

Subunit structure

Forms homodimers via its C-terminal domain (CTD) in the presence of metal cations (Probable).

Interacts with MamB via their CTD (PubMed:22007638, PubMed:29243866) (Probable). Isolated CTD forms homodimers (PubMed:24658343, PubMed:24819161, PubMed:27550551, PubMed:30811856).

3 Publications 6 Publications

Рисунок 2.10 – Основна інформація про білок: процесинг

Недавнім оновленням стала можливість отримати візуалізацію 3D-структури білка прямо на сторінці баз даних UniProt, не переходячи при цьому на сторонні ресурси (рис. 2.11). Структура відображається лише за наявності даних, підтверджених експериментально. За відсутності таких даних необхідно використовувати інші ресурси (посилання на які також представлені на сторінці результату пошуку обраного білка) для побудови такої структури за відомою амінокислотною послідовністю.

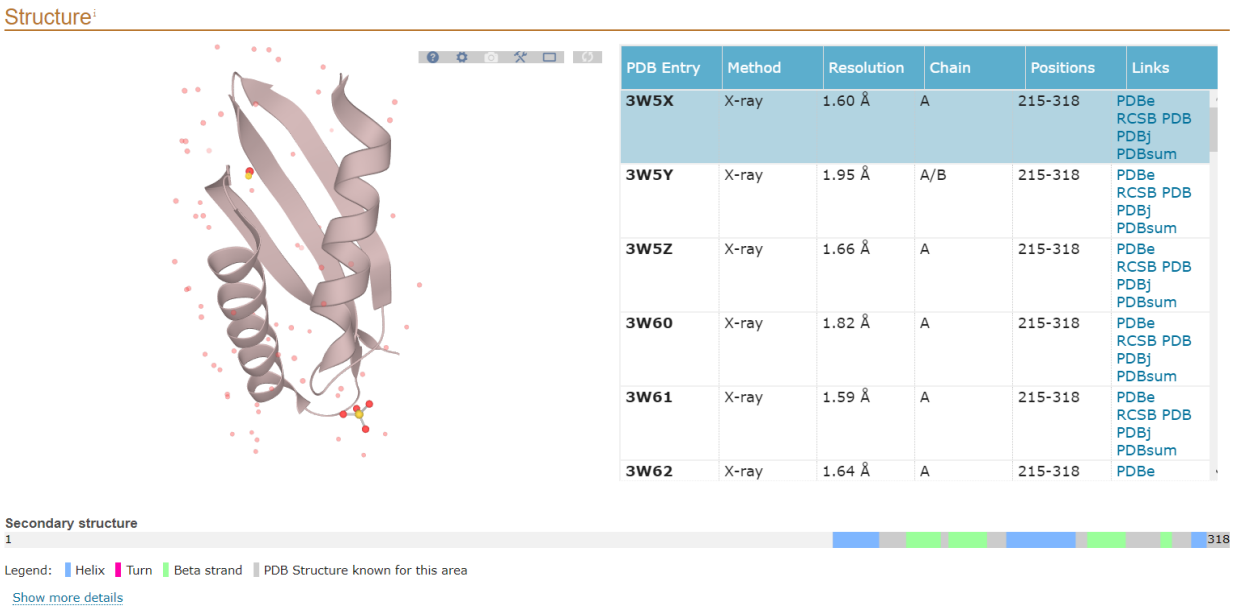


Рисунок 2.11 – Інформація про білок: структура

Окрім інформації, наведеної вище, у базах даних UniProt є додаткові функції, що дозволяють отримувати більше даних про амінокислотну послідовність:

- BLAST – програма, за допомогою якої можна вирівняти послідовність із усіма наявними імовірними гомологами у базі даних;
- Clustal Omega – програма для множинного вирівнювання послідовностей;

- Retrieve/ID mapping – програма, що дозволяє змінювати формат послідовності чи проводити пошук записів про той чи інший білок;
- Peptide search – програма, що дозволяє проводити пошук білка за відомою амінокислотною послідовністю.

Нижче на рис. 2.12 та рис. 2.13 представлені результати вирівнювань досліджуваного білка за допомогою ресурсів BLAST та Clustal Omega, які вбудовані у UniProt. При вирівнювання за допомогою програми BLAST можливо вибрати лише одну амінокислотну послідовність. Результатом є вирівнювання, розташоване у порядку спадання відсотка ідентичності між послідовностями. Різні діапазони значень ідентичності представлені різними кольорами.

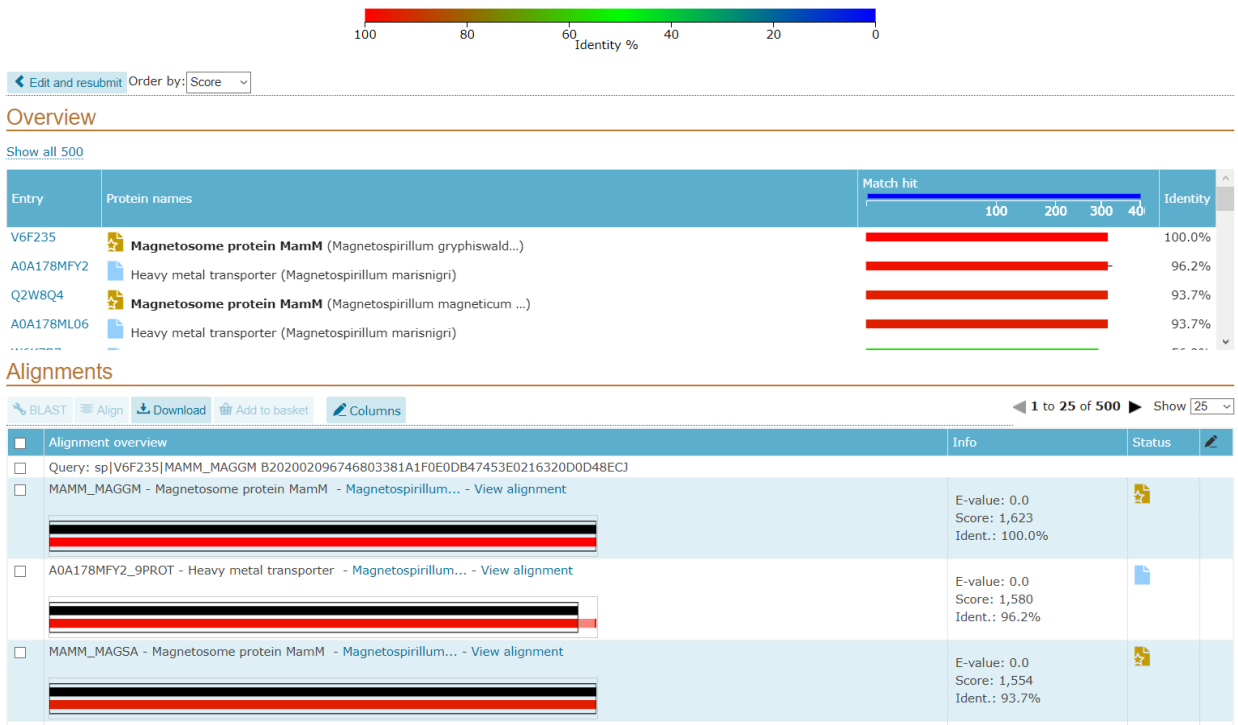


Рисунок 2.12 – Результат вирівнювання за допомогою програми BLAST, вбудованої у базу даних UniProt

Для множинного вирівнюваннння необхідно вибрати три або більше

послідовностей. Ресурс дозволяє виділяти різними кольорами певні властивості білків (активні центри, сайти, мутації тощо) чи окремих амінокислот (заряд, розмір, гідрофільність тощо).

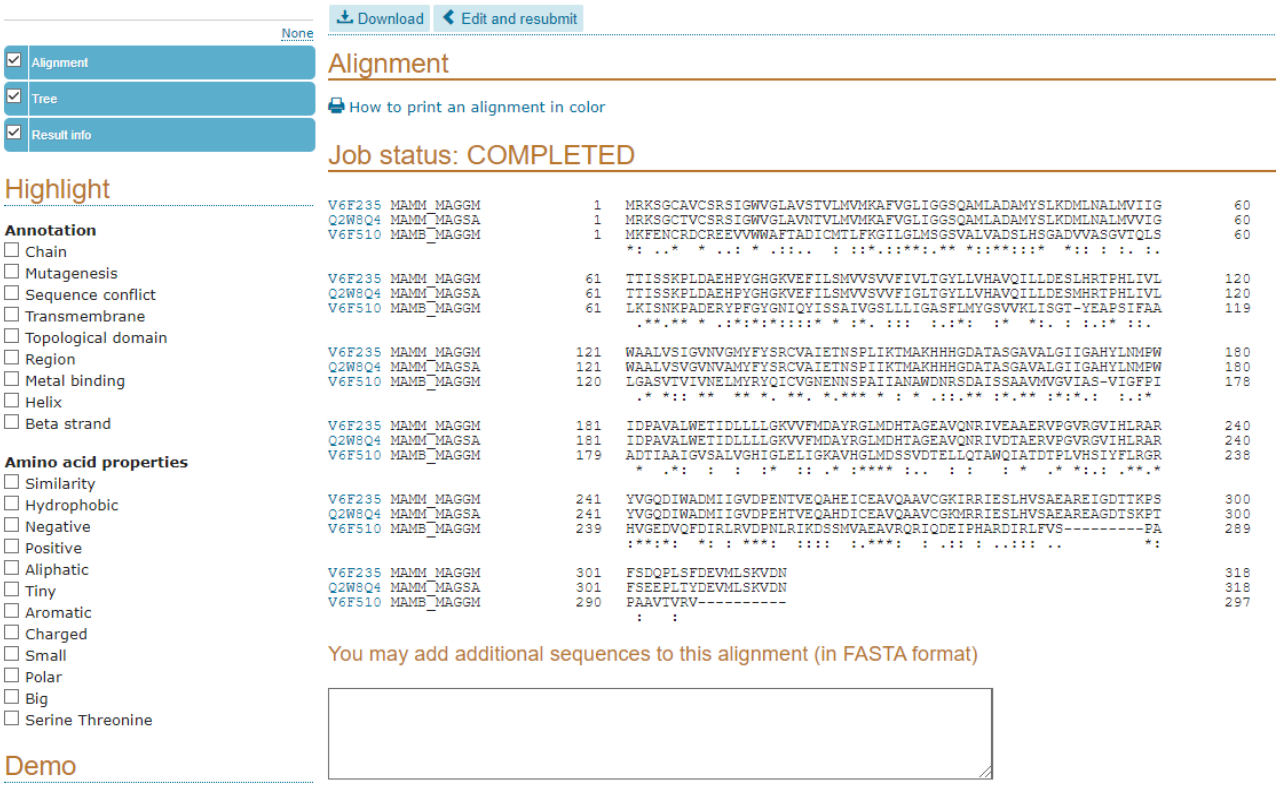


Рисунок 2.13 – Результат вирівнювання білкових послідовностей за допомогою програми Clustal Omega

Отже, база даних UniProt володіє широким різноманіттям інформації про велику кількість білків та білкових послідовностей. Уся інформація синхронізована із літературними джерелами та іншими базами даних, про що свідчать відповідні посилання. Додаткові можливості (наприклад, вирівнювання послідовностей) дозволяють отримати ще більше інформації про той чи інший білок, при цьому залишаючись на сторінці основної бази даних UniProt.

2.2 База даних PROSITE

У деяких випадках послідовність невідомого білка занадто віддалена з будь-яким білком відомої структури, щоб виявити його схожість шляхом загального вирівнювання послідовностей, але його можна ідентифікувати за появою у його послідовності певного кластера залишків, який відомий як патерн, мотив або домен. Ці мотиви виникають через особливі вимоги до структури конкретних областей білка, які можуть бути важливими, наприклад, щодо їх зв'язуючих властивостей або для їх ферментативної активності. Ці вимоги накладають дуже жорсткі обмеження за розміром, але при цьому є важливими частинами послідовності білка.

PROSITE – це метод визначення функції білків, перекладених з геномних або кДНК-послідовностей. Він складається з бази даних біологічно значущих сайтів та моделей, сформованих таким чином, що за допомогою відповідних обчислювальних інструментів можна швидко та надійно визначити, до якої відомої родини білків (якщо така є) належить нова послідовність.

Існує ряд сімейств білків, а також функціональних або структурних доменів, які неможливо виявити за допомогою шаблонів через надзвичайну розбіжність послідовностей; в такому випадку використання методик на основі вагових матриць (також відомих як профілі) дозволяє виявити такі білки чи домени. В даний час більшість нових записів PROSITE зосереджені навколо профілів і розробляються співробітниками PROSITE Швейцарського інституту біоінформатики SIB у Женеві та Лозанні.

PROSITE – це ресурс для ідентифікації та анотації певних мотивів та доменів у білкових послідовностях. Ці ділянки ідентифікуються за допомогою двох типів так званих «записів»: узагальнені профілі (вагові матриці), що описують сімейства білків та модульні білкові домени та шаблони (регулярні

вирази), що описують мотиви короткої послідовності, що часто відповідають функціонально або структурно важливим залишкам. Підписи PROSITE пов'язані з правилами анотації або ProRules, які визначають анотації послідовностей білків (наприклад, активних сайтів та залишки, що зв'язують ліганд) та умови, при яких вони застосовуються. ProRules використовує для анотації сімейства білків, доменів та особливостей послідовностей в UniProtKB/Swiss-Prot, відредагований вручну розділ UniProt KnowledgeBase. На даний час він надає анотацію для більш ніж 75% з 1054 доменів, які там можна знайти. Частина інформації, що зберігається в ProRules (наприклад, активні сайти та сайти зв'язування, дисульфідні зв'язки), також доступні користувачеві через ScanProsite.

PROSITE надає велику кількість документації для кожного запису, включаючи інформацію про номенклатуру, функції, особливості послідовності, тривимірні структури (структури), білкові архітектури, в яких знайдено відповідний запис, його таксономічний розподіл та важливі літературні посилання. Записи на PROSITE також доступні через InterPro, інтегровану базу даних білкових записів, яка використовується для класифікації та анотації білків та геномів. З моменту останнього звіту у випуску журналу Nucleic Acid Research (NAR), БД PROSITE збільшила кількість доступних записів до 1308 шаблонів та 1039 профілів, які пов’язані з 1041 записами ProRules та 1650 документами.

БД PROSITE розроблялася паралельно з Swiss-Prot, і обидві бази даних отримували користь одна від одної. Синтаксис шаблону PROSITE адаптований для коротких добре збережених ділянок. Такі ділянки, як правило, є активними центрами ферментів, місцями приєднання до простетичних груп (гема, піридоксальфосфату, біотину тощо), амінокислоти, що зв'язують іони металів, цистеїни, що беруть участь у дисульфідних зв’язках, або області, що беруть участь у зв'язуванні молекули. Але цей синтаксис дуже чутливий до будь-якої послідовності «виключення», чи то через розбіжність, чи через помилку послідовності. Таким чином, шаблони не пристосовані для ідентифікації менш

збережених ділянок або цілих доменів. У 1994 році Філіп Бюхер представив у PROSITE «узагальнені профілі» як нові дескриптори мотивів. Усі профільні методи – це статистичні описи множинного вирівнювання послідовностей. Оскільки синтаксис «узагальненого профілю» дуже схожий на профіль HMM (Hidden Markov Models), майже всі результати «узагальненого профілю» можуть бути зіставлені з параметрами HMM, які використовуються пакетом програм для аналізу послідовностей HMMER. Наразі майже всі нові записи PROSITE є профілями. З часу свого створення PROSITE надав обширну документацію та детальну анотацію доменів, сімейств та функціональних сайтів. Ця інформація в основному зберігається у вільному доступі та використовується біологами, які приймали рішення щодо функції досліджуваного білка відповідно до даних PROSITE. Але зі швидким зростанням баз даних послідовностей протягом останніх 10 років зросла потреба у надійному інструменті, який міг би генерувати автоматично функціонально точну анотацію у стандартному форматі. Тому 2005 році було вирішено згрупувати деяку функціональну інформацію, що зберігається в PROSITE, у базі даних правил ProRules.

2.2.1. Оновлення у базі даних PROSITE

PROSITE можна переглядати за таксономічними областями, за описом ProRules, за кількістю позитивних збігів або за відповідними білками. Була реорганізована презентація даних, які тепер згруповані в п'ять різних розділів (ScanProsite (Аннотація), ProRules (Автоматизована аннотація), Documents (Документи), Loading (Завантаження) та Reference (Посилання). Створено новий розділ ProRule, який дозволяє візуалізувати різні правила, які використовуються для генерування анотацій на веб-сторінці «ScanProsite».

Дуже простий синтаксис дозволяє користувачеві визначити форму, колір, розмір та назву одного або декількох доменів. Також можуть бути відмічені

конкретні залишки та діапазони послідовностей.

Ще одним оновленням є інструмент «MyDomains», який генерує зображення домену для представлення білкових архітектур. Для даного білка користувач може ввести у веб-форму розмір білка, положення доменів, їх назви та для кожного домену колір та форму потрібного зображення. Веб-форма повертає зображення у форматі Portable Network Graphics (PNG) (рис. 2.14).

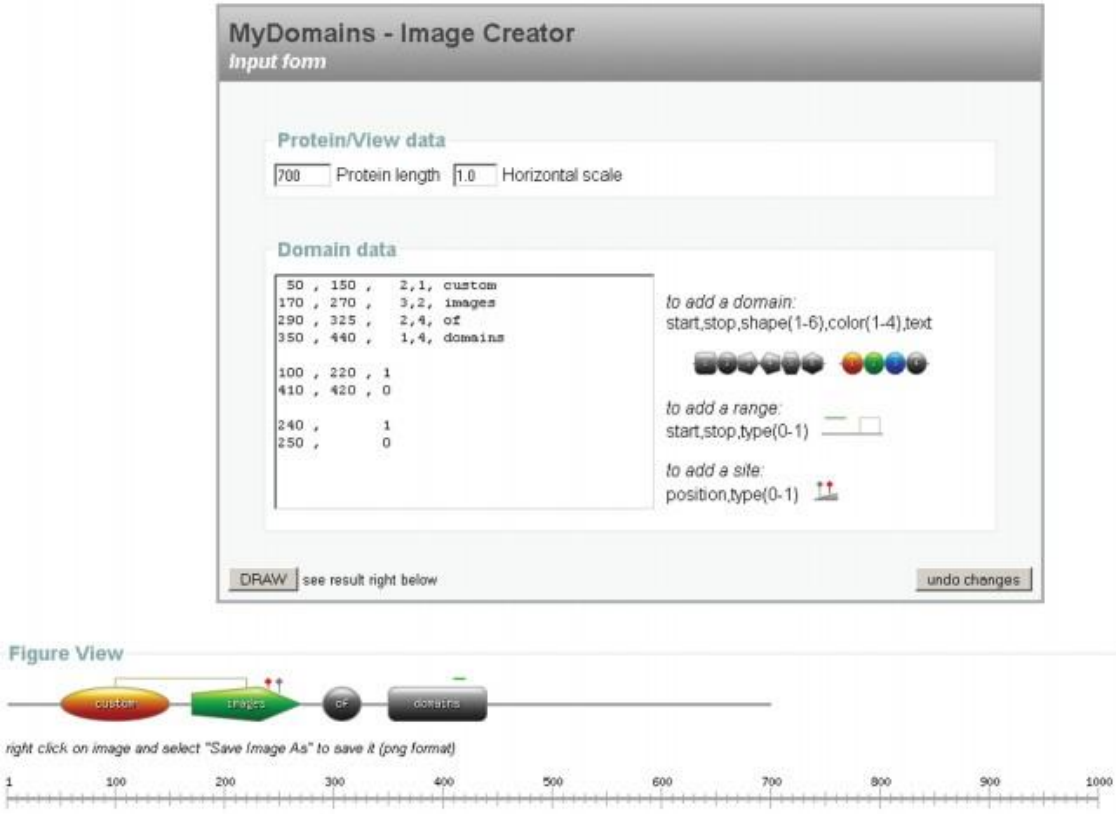


Рисунок 2.14 – Веб-форма інструменту для створення зображень PROSITE «MyDomains»

2.2.2.Основні можливості бази даних PROSITE

У базі даних PROSITE міститься інформація про домени, білкові родини та функціональні сайти. З його допомогою було проведено аналіз обраного білка

(MamA магнітотаксисної бактерії *Magnetospirillum gryphiswaldense* MSR-1) на наявність таких сайтів. На стартовій сторінці ресурсу міститься два поля пошуку, а також додаткові можливості (рис. 2.15).

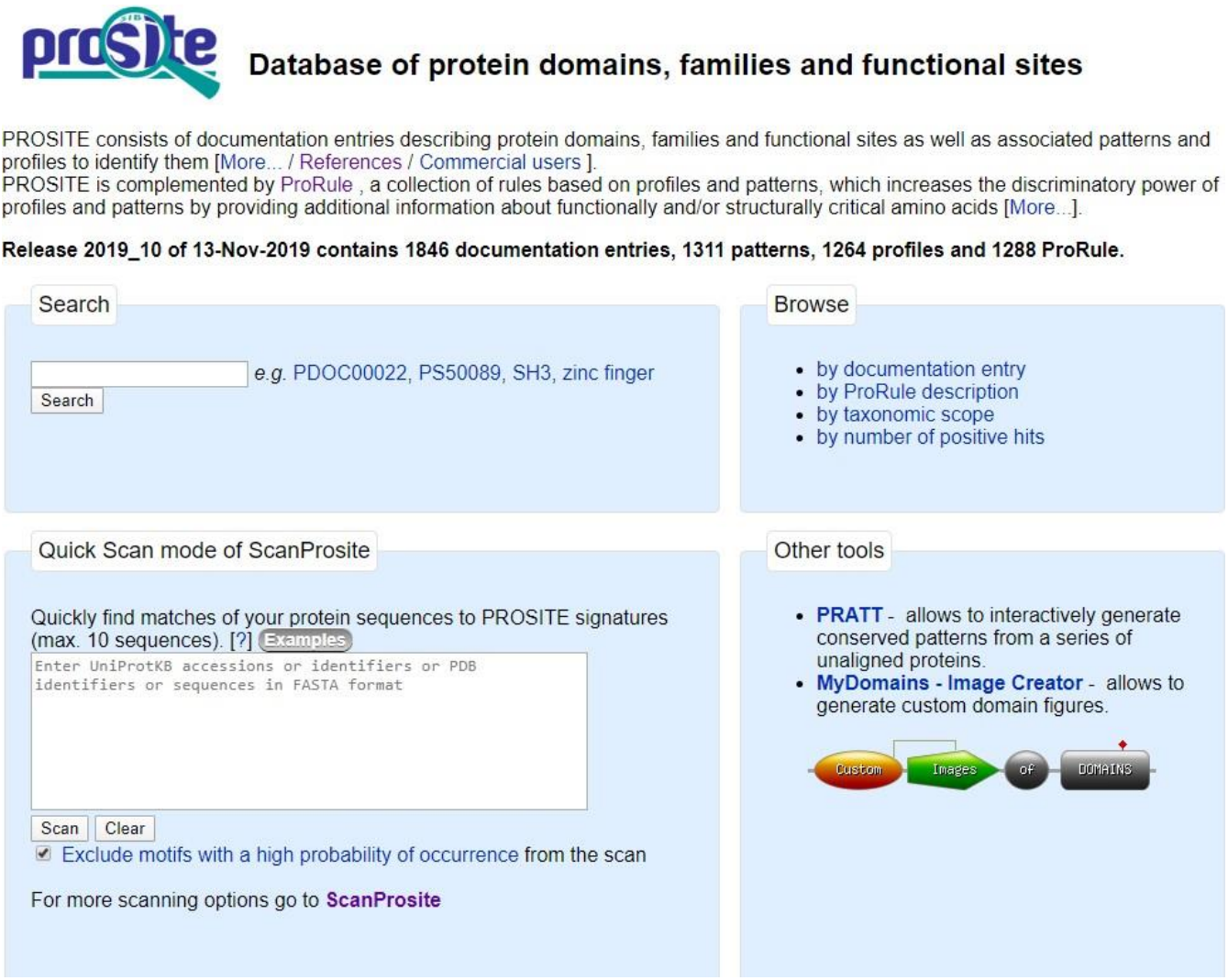


Рисунок 2.15 – Стартова сторінка бази даних PROSITE

Пошук за доменами можливо здійснювати за допомогою логічних операторів: «і», що включає обидва зазначені домени, «або», що включає як наявність обох доменів разом, так і кожного окремо, а також «але не», що виключає пошук певного домена (рис.2.16).

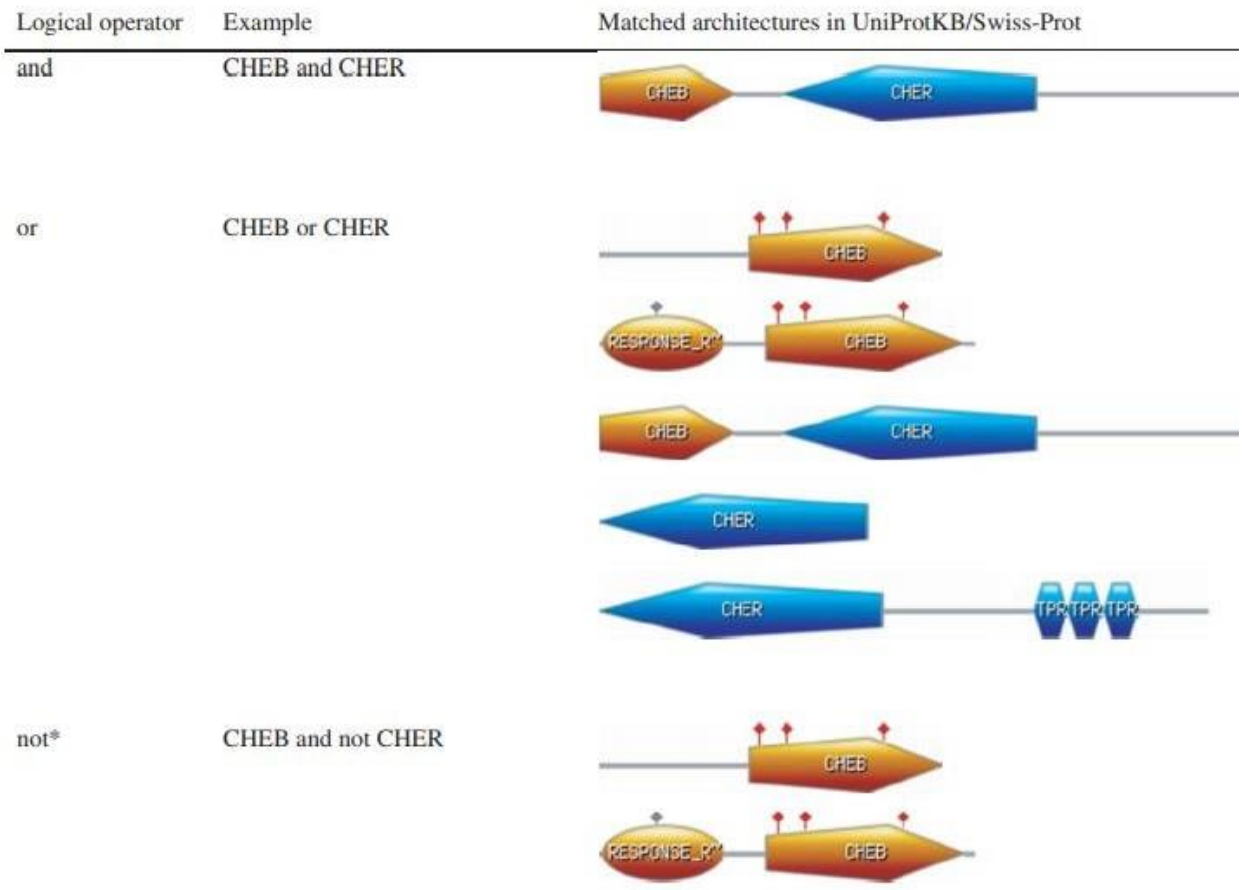


Рисунок 2.16 – Використання логічних операторів при пошуку

Було здійснено аналіз білка MamA магнітотаксисної бактерії *Magnetospirillum gryphiswaldense* MSR-1. База даних UniProt містить посилання на PROSITE, яке й було використано для доступу до запису про білок у базі даних PROSITE. Результатом стало схематичне зображення домену, виявленого у даному білку та його опис (рис. 2.17).

Рисунок 2.17 – Результат пошуку доменів білка MamA магнітотаксисної бактерії *Magnetospirillum gryphiswaldense* MSR-1 у PROSITE

Запитання до розділу 2

1. Назвіть найбільш відомі БД білків.
2. Які бази даних об’єднує в собі БД UniProt?
3. Яка БД містить інформацію про послідовності протеїнів та дані щодо їх функцій, отримані з проектів сиквенсу геномів?
4. Назвіть БД, яка містить просторові структури білків, визначені дослідним шляхом, еволюційну історію і взаємозв’язок між макромолекулами.
5. Яка БД містить інформацію щодо номенклатури ферментів, і описує всі типи білків, яким присвоєно номер EC (Enzyme Commission)?
6. По яких критеріях реалізовано пошук в БД ENZYME?
7. В якому розділі UniProt зберігається інформація, взята з наукових публікацій?
8. Дати визначення пан-протеому.

9. Які програми забезпечують перегляд білкової структури в UniProt?
10. Назвіть додаткові функції в базі даних UniProt, що дозволяють отримувати більше даних про амінокислотну послідовність.
11. З якою метою створено БД PROSITE?
12. З якою БД паралельно розроблявся PROSITE?
13. Назвіть розділи, в які згруповано дані в PROSITE.
14. За якими параметрами можна переглядати інформацію в PROSITE?
15. Назвіть інструмент в PROSITE, який генерує зображення домену для представлення білкових архітектур.

Література до розділу 2

1. Binder, J.X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S.I., Schneider, R. and Jensen, L.J. COMPARTMENTS: unification and visualization of protein subcellular localization evidence. Database, 2014

2. Breuza,L., Poux,S., Estreicher,A., Famiglietti,M.L., Magrane,M., Tognolli,M., Bridge,A., Baratin,D., Redaschi,N. and UniProt Consortium. (2016) The UniProtKB guide to the human proteome. Database, 2016

3. Chen, C., Huang, H., & Wu, C. H. Protein Bioinformatics Databases and Resources. Methods in Molecular Biology, 3–39, 2017

4. Chen,C., Huang,H., Mazumder,R., Natale,D.A., McGarvey,P.B., Zhang,J., Polson,S.W., Wang,Y., Wu,C.H. and Consortium,UniProt Computational clustering for viral reference proteomes. Bioinformatics, 32, 2041–2043, 2016

5. Huang, H., Hu, Z., Suzek, B.E., & Wu, C.H. The PIR integrated protein databases and data retrieval system. Data Science Journal, 3, 163-174, 2004

6. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J.A. The 20 years of PROSITE. Nucleic Acids Res., 36, D245–D249, 2008

7. Medini,D., Donati,C., Tettelin,H., Masignani,V. and Rappuoli,R. The microbial pan-genome. Curr. Opin. Genet. Dev., 15, 589–594, 2005

8. Poux,S., Magrane,M., Arighi,C.N., Bridge,A., O'Donovan,C., Laiho,K. and

UniProt,C. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database, 2014

9. PROSITE. URL: https://prosite.expasy.org/

10. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I. New and continuing developments at PROSITE. Nucleic Acids Res. 2012

11. Sigrist,C.J.A., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res., 38, D161–D166, 2010

12. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B., Wu,C.H., UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics, 31, 926–932, 2015

13. The UniProt Consortium, UniProt: the universal protein knowledgebase, *Nucleic Acids Research*, Volume 46, Issue 5, 2018

14. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic acids research*, *47*(D1), D506-D515, 2018

15. UniProt KnowledgeBase. URL: http://www.uniprot.org

16. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Giron,C.G. ´ et al. Ensembl 2018. Nucleic Acids Res., 46, D754–D761, 2018

Розділ 3. БД СПЕЦІАЛІЗОВАНИХ БІОІНФОРМАТИЧНИХ РЕСУРСІВ

Успіхи у розшифровці геному людини, швидкий розвиток молекулярної біології та генетики, створення сотень баз даних геномів, білків, метаболомів, окремих організмів, хвороб, експресії генів, сигнальних молекул, тощо, сприяли появі нових наукових напрямків таких як геноміка, протеоміка, метаболоміка, фармакогеноміка та привели до виникнення абсолютно нового підходу до лікування захворювань, такого як генна терапія. Принципова відмінність нового способу лікування від традиційних полягає в тому, що він спрямований на усунення першопричини захворювання, а не її наслідків. На сучасному етапі генну терапію можна визначити як лікування спадкових і неспадкових захворювань шляхом введення генів у клітини пацієнтів з метою спрямованої зміни генних дефектів або надання клітинам нових функцій.

Генна терапія досягла значних успіхів у боротьбі з пухлинними захворюваннями. До теперішнього моменту розроблені кілька основних підходів. Перш за все, це нормалізація роботи онкогенів і супресорів пухлин. Не менш перспективним видається й інший підхід, пов'язаний з навчанням імунної системи розпізнавати антигени ракових клітин. На цьому принципі засновано створення протипухлинних вакцин. Відчутні результати отримані в області нейродегенеративних захворювань, таких як хвороба Паркінсона, хорея Гентингтона, в лікуванні ВІЛ-інфікованих хворих, в кардіології, а також в ряді інших захворювань.

Досягнення генної терапії, яка є розділом геноміки, пов’язаним з секвенуванням і аналізом геному конкретної людини, сприяли розвитку персоналізованої медицини, яка використовує інформацію персональної геноміки при виборі медичних процедур, необхідних для конкретної людини з урахуванням індивідуальних особливостей до яких крім генетичних маркерів відносяться эпігенетичні, транскриптомні, протеомні, метаболомні і метагеномні маркери, а також сукупність варіативних фенотипових ознак як всього організму пацієнта, так і його окремих тканин або клітин.

В сфері персоналізованої медицини на даний час є доступними значна кількість проектів і вже доступних послуг:

• Генографічний проект - проект Національного Географічного Товариства і IBM, які збирають зразки ДНК для відтворення моделей історичних міграцій людини. Він був прийнятий в 2005 р (з 500,000 учасників станом на грудень 2012 року), що допомогло створити доступну для споживача (DTC) тестову генетичну промисловість.

• Персональний Геномний Проект (PGP) - довгострокова велика компанія, заснована в Гарвардської Медичної Школі, яка поставила перед собою мету секвенування і публікації готових геномів і медичних документів 100,000 добровольців, щоб направити дослідження на персональну геноміку і персональну медицину.

• SNPedia - це вікі, що збирає і поширює інформацію про наслідки варіацій ДНК, і через відповідну програму Promethease кожен, хто отримав ДНК дані про себе (від будь-якої компанії) може отримати вільний, незалежний звіт, що містить оцінку ризиків і пов'язану з нею інформацію.

• deCODEme.com бере 1100 $ для проведення генотипування близько 1 млн SNP, і надає оцінки ризику для 47 хвороб і аналіз родоводу.

• Existence Genetics надає послуги генетичного тестування через органи охорони здоров’я і організації здорового способу життя за вартістю, починаючи від $ 299. Ця компанія забезпечує тестування понад 1200 поширених і рідкісних захворювань і характерних для них властивостей, включаючи захворювання серця, рак, аутоімунні захворювання, ожиріння, нутрігеноміку, фармакогеноміку, фітнес і спортивні показники. Existence Genetics забезпечує центри генетичного тестування результатів в фітнесі та спорті для Equinox Fitness.

• Navigenics почали пропонувати SNP на основі геномної оцінки ризику за станом на квітень 2008 року. Navigenics підкреслює роль лікарів в розшифровці генетичних і медичних результатів. Генотипи Affymetrix Genome-Wide Human SNP Array 6.0 складаються з 900, 000 SNP.

• Pathway Genomics проводять аналіз понад 100 генетичних маркерів для виявлення ризику генетичних захворювань, таких як меланома, рак передміхурової залози і ревматоїдного артриту.

• 23andMe продають комплекти для SNP генотипування поштою. Інформація зберігається в профілі користувача і використовується для оцінки ризику 178 генетичних захворювань пацієнта і аналізування родоводу. 23andMe використовує масиви ДНК виробництва фірми Illumina.

Безумовно розвиток таких проектів та послуг персоналізованої медицини неможливий без використання та розвитку як вищеперерахованих, так і БД спеціалізованих біоінформатичних ресурсів таких як: баз даних метаболічних шляхів, баз даних сполук, баз даних спектрів, баз даних захворювань/фізіології, баз даних мутацій, комплексних, специфічних для організму баз даних метаболому, баз даних SNP, баз даних експресії генів, тощо.

Розглянемо деякі з цих БД більш докладно.

3.1 Бази даних метаболічних шляхів

Метаболоміка є корисною в різних сферах, включаючи виявлення або розробку лікарських засобів, клінічну токсикологію, ідентифікацію біомаркерів, дослідження харчових продуктів та кількісну фенотипізацію рослин або мікроорганізмів. Вивчення можливостей та принципів роботи із базами даних метаболому людини є важливою задачею, оскільки метаболоміка може бути корисною в інтерпретації багатьох складних біологічних процесів.

Метаболом – це повний набір низькомолекулярних речовин – метаболітів в тому чи іншому біологічному зразку в конкретний момент часу. Під низькомолекулярними речовинами зазвичай розуміють молекули з молекулярною масою до 1500 Да. В якості зразка можуть бути клітини, тканини, екстракти тканин, органи, біологічні рідини або цілий організм.

Низькомолекулярні речовини метаболому можуть утворюватись в організмі природним шляхом (амінокислоти, органічні кислоти, нуклеїнові кислоти, вітаміни, пігменти, цукри та ін.) або надходити ззовні (ліки, харчові добавки, токсини та ін.). Тому розділяють ендогенний та екзогенний метаболоми.

Оскільки метаболоміка поєднує молекулярну біологію з хімією та фізіологією, існує необхідність не одного типу баз даних, а широкого спектру електронних ресурсів.

В даний час існує щонайменше 5 типів баз даних, що використовуються в дослідженнях метаболоміки. До них належать: бази даних метаболічних шляхів, бази даних сполук, бази даних спектрів, бази даних хвороб/фізіології, комплексні бази даних, пов’язані з метаболомом окремих організмів, тощо.

База даних метаболому людини або HMDB – це найбільша та найповніша база даних метаболому, специфічного для організму. Вона містить спекстроскопічну, кількісну, аналітичну, та молекулярно-масштабну інформацію про метаболіти людини, пов’язані з ними ферменти або транспортери, їх чисельність та хвороботворні властивості. На цей час HMDB вважається основним стандартним ресурсом для вивчення метаболізму людини. Протягом останнього десятиліття HMDB продовжувала рости і розвиватися у відповідь на нові потреби дослідників метаболоміки.

Бази даних метаболічних шляхів містять ретельно проілюстровані, гіперпов’язані метаболічні шляхи з інформацією про метаболіти для широкого кола організмів.

KEGG (Kyoto Encyclopedia of Genes and Genomes) – одна із найповніших і широко використовуваних баз даних, що містять метаболічні шляхи найрізноманітніших організмів (> 700). Ці шляхи гіперпов'язані з інформацією про метаболіт та білок / фермент. В даний час KEGG містить > 15000 сполук (тварин, рослин і бактерій), 7742 лікарських засобів (включаючи різні сольові форми та носії лікарських засобів) та майже 11 000 гліканових структур.

‘Cyc’ бази даних:

MetaCyc – база даних не надлишкових експериментально з'ясованих метаболічних шляхів, що містить інформацію про шляхи, що беруть участь як у первинному, так і у вторинному метаболізмі, а також пов'язані сполуки, ферменти та гени.

HumanCyc – біоінформатична база даних, що містить інформацію про метаболічні шляхи та геном людини. База даних включає інформацію про гени, їх продукти та метаболічні реакції і шляхи, які вони каталізують.

BioCyc – колекція із 371 баз даних геномів/метаболічних шляхів. Кожна база даних із колекції описує геном та метаболічні шляхи одного окремого організму.

Reactome – це база даних, що містить інфомацію про біологічні шляхи, включаючи метаболічні шляхи, а також білковий транспорт та сигнальні шляхи. Reactom має данні про шляхи більше, ніж 20 різноманітних організмів, але основним організмом, що є людина.

WikiPathways – це відкрита платформа для збору та розповсюдження моделей біологічних шляхів для візуалізації та аналізу даних.

BiGG – це база даних метаболічної реконструкції метаболізму людини, призначена для моделювання біологічної системи та моделювання балансу метаболічного потоку.

MetaboLights – база даних експериментів метаболоміки та отриманої інформації. База даних є міжвидовою та охоплює структури метаболітів та їх еталонні спектри, а також біологічні ролі, розташування, концентрації та експериментальні дані з метаболічних експериментів.

HMDB (Human Metabolome Database) – вільнодоступна електронна база даних, що містить детальну інформацію про метаболіти малих молекул, виявлених (та підтверджених експериментально) в організмі людини. База даних містить 3 види даних: хімічні, клінічні та дані молекулярної біології/біохімії. Кожен запис у MetaboCard має більше 100 полів даних, при

чому 2/3 інформації присвячено хімічним або клінічним даним, а 1/3 – ферментативним або біохімічним даним. Багато полів даних мають гіперпосилання на інші бази даних (KEGG, PubChem, MetaCyc, ChEBI, PDB, Swiss-Prot и GenBank) та різноманітні аплети для перегляду структур та шляхів.

3.2 Бази даних метаболома людини Human Metabolome Database (HMDB)

Перша версія HMDB була випущена 1 січня 2007 року, після чого дві наступні версії 1 січня 2009 року (версія 2.0), 1 серпня 2009 (версія 2.5), 18 вересня 2012 року (версія 3.0), 1 січня 2013 року (версія 3.5), 2017 (версія 4.0).

На рис. 3.1 зображено таблицю порівняння чотирьох версій бази данних за вмістом доступної інформації.

Comparison between the coverage in HMDB 1.0, 2.0, 3.0 and HMDB 4.0

Category	HMDB 1.0	HMDB 2.0	HMDB 3.0	HMDB 4.0
Total number of metabolites	2180	6408	40 153	114 100
Number of detected & quantified metabolites	883	4413	16 714	18 557
Number of detected, not quantified metabolites	1297	1995	2798	3271
Number of expected metabolites	0	0	20 641	82 274
Number of predicted metabolites*	0	0	0	9548
Number of unique synonyms	27 700	43 882	199 668	1 231 398
Number of cmpds with expt. MS/MS spectra	390	799	1249	2265
Number of cmpds with expt. GC/MS spectra	0	279	1220	2544
Number of cmpds with expt. NMR spectra	385	792	1054	1494
Number of cmpds with pred. MS/MS spectra*	0	0	0	98 601
Number of cmpds with pred. GC/MS spectra*	0	0	0	26 880
Number of experimental NMR spectra	765	1580	2032	3840
Number of experimental MS/MS spectra	1180	2397	5776	22 198
Number of experimental GC/MS spectra	0	279	1763	7418
Number of predicted MS/MS spectra*	0	0	0	279 972
Number of predicted GC/MS spectra*	0	0	0	38 277

Рисунок 3.1 – Порівняння вмісту різних типів даних у різних версіях HMDB

Цей проект підтримується Канадськими інститутами медичних досліджень, Канадським фондом інновацій та Інноваційним центром метаболоміки (TMIC), що фінансується на національному рівні науково-дослідницьким центром, який підтримує широкий спектр сучасних метаболомних досліджень. Компанія TMIC фінансується за рахунок Genome Canada, Genome Alberta та Genome British Columbia, некомерційною організацією, яка очолює канадську національну стратегію в області геноміки з фінансуванням федерального уряду на суму $ 900 млн.

В базі даних є багато посилань на інші бази даних, такі як KEGG, PubChem, MetaCyc, ChEBI, PDB, UniProt, GenBank, DrugBank та інші. А також вона включає 3 додаткові бази даних T3DB, SMPDB і FooDB.

3.2.1 Основні інструменти пошуку в БД HMDB

Human Metabolome Database (рис. 3.2) – це багатоцільова база даних біоінформатики, хімічної інформатики та медичної інформатики, яка має сильний акцент на кількісну, аналітичну або молекулярно-масштабну інформацію про метаболіти, пов’язані з ними ферменти або транспортери та їх властивості, асоційовані з хворобами.

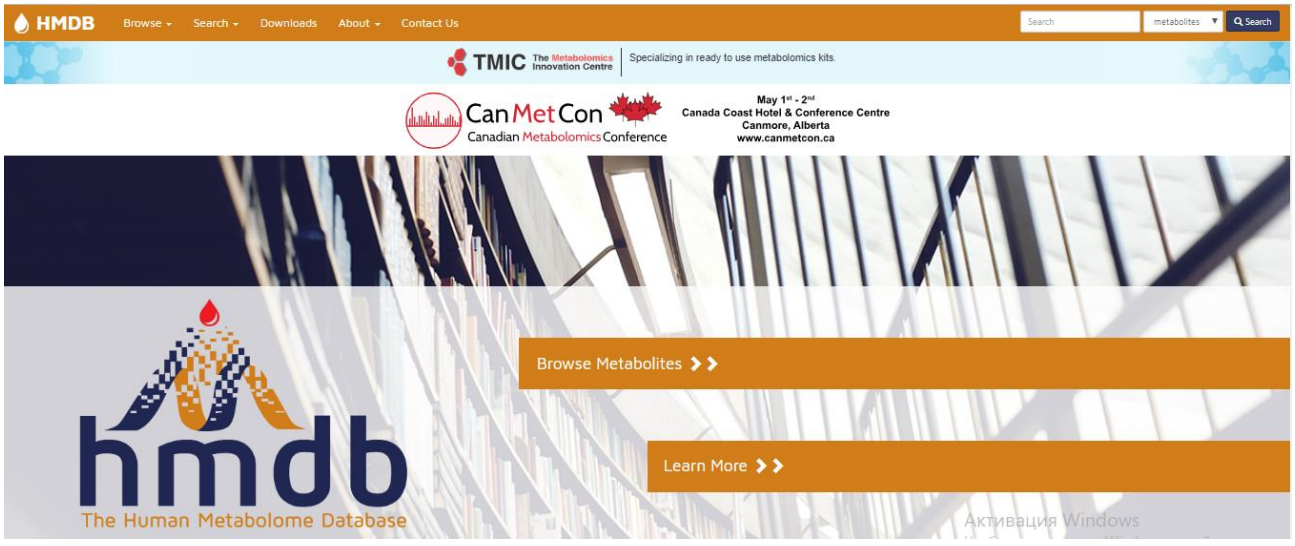


Рисунок 3.2 – Вигляд сайту HMDB

У багатьох відношеннях HMDB поєднує численні дані молекулярної біології, зазвичай знайдених у базах даних послідовностей, таких як SwissProt і UniProt, з настільки ж багатими даними, що містяться в KEGG (про метаболізм) і OMMBID (про клінічні умови). Вона також містить великий масив незалежних експериментальних даних, включаючи спектри ЯМР, спектри MS, дані розчинності та підтверджені концентрації метаболітів, для доповнення літературних даних.

Різноманітність типів даних, кількість експериментальних даних та необхідна кількість знань в предметній області зробили створення HMDB складною та трудомісткою. При створенні HMDB для узагальнення, перевірки та підтвердження комплексної колекції даних було проведено індивідуальний пошук та порівняння більше двох десятків підручників, кількох тисяч статей, майже 30 різних електронних баз даних, і принаймні 20 веб-програм. Крім того, було зібрано понад 2100 спектрів ЯМР та MS, здійснено 160 експериментальних визначень концентрацій, проведено 75 органічних синтезів і виконано сотні високоефективних рідинних хроматографій (ВЕРХ).

Команда авторів та анотаторів HMDB включала трьох хіміків-органіків, шість спектроскопістів ЯМР, п'ять мас-спектроскопістів, двох спеціалістів з розділення, три лікаря та 14 біоінформатиків з подвійною освітою у галузі обчислювальної техніки та молекулярної біології чи хімії.

HMDB містить більше 2180 записів метаболітів людини, які пов'язані з більш ніж 27 700 різними синонімами. Ці метаболіти додатково пов'язані з 115 метаболічними шляхами, 2080 окремими ферментами, 110 000 SNP, а також 862 метаболічними захворюваннями (генетичними та набутими).

Більш ніж 400 сполук також пов'язані з експериментально отриманими «еталонними» ЯМР і MS спектрами. Дані по концентраціям (нормальні та аномальні значення) для плазми, сечі та / або інших біологічних рідин також представлені для 883 сполук. Вся база даних, включаючи текст, дані про

послідовності, структуру та зображення, займає майже 18 ГБ даних, більшість з яких можна вільно завантажувати.

HMDB є повністю доступною для пошуку завдяки великій кількості вбудованих інструментів для перегляду, сортування та вилучення метаболітів, концентрацій біологічних рідин, ферментів, генів, ЯМР або MS спектрів та інформації про захворювання. Детальні інструкції щодо того, де знайти та як використовувати ці засоби перегляду та пошуку, надаються на домашній сторінці HMDB. Як і будь-яка база даних, HMDB підтримує стандартні текстові запити через текстове поле пошуку, розташоване у верхній частині кожної сторінки. Щоб полегшити перегляд даних, HMDB розділена на зведені таблиці, які в свою чергу, пов’язані з «метабокартами» (рис. 3.3) - за аналогією з дуже успішною концепцією DrugCards, яка використовується в DrugBank. Усі зведені таблиці HMDB можна швидко переглядати, сортувати або форматувати у спосіб, подібний до того, як можуть бути переглянуті тези на PubMed.

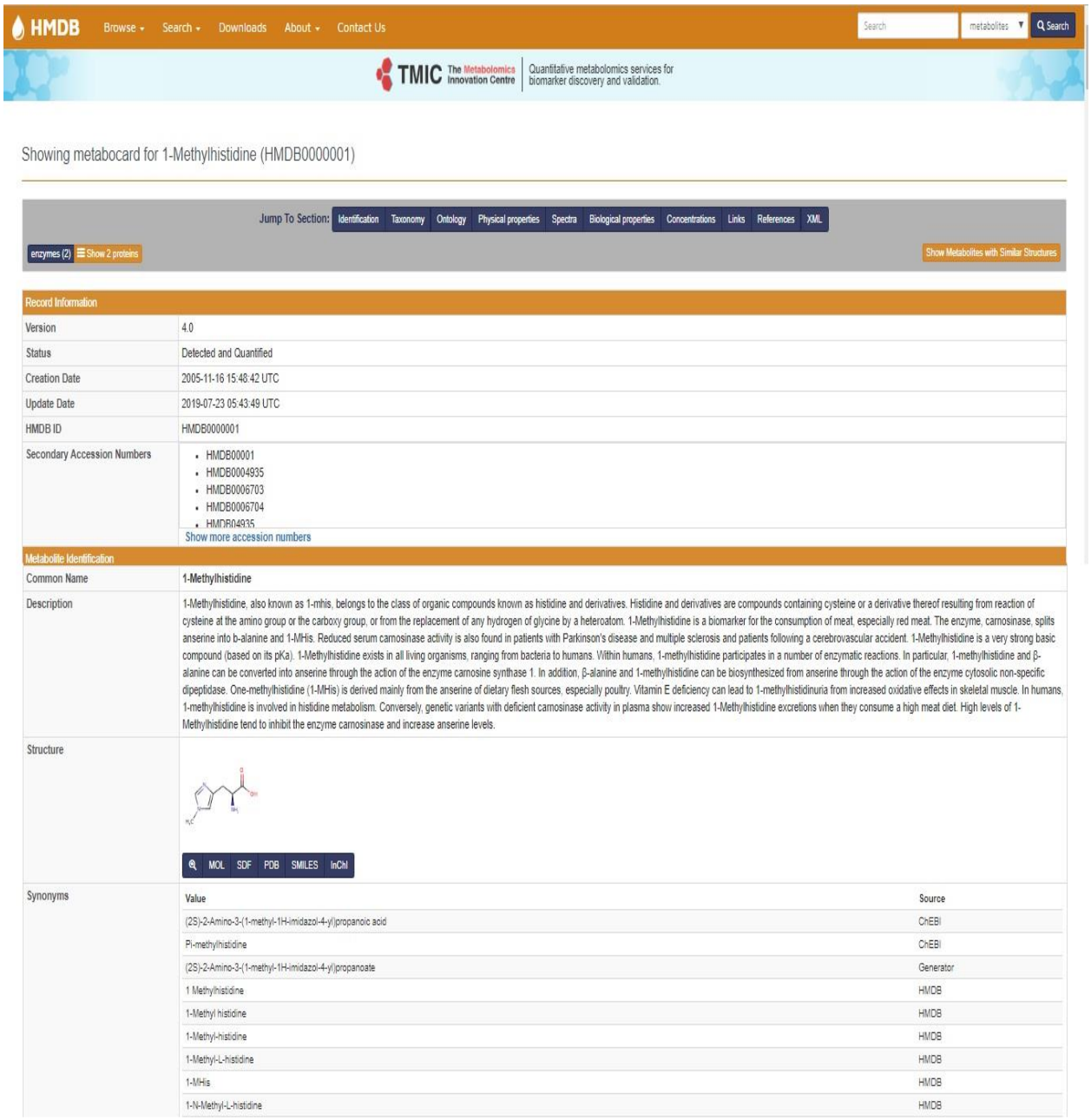


Рисунок 3.3 – Метабокарта метилгістидину у HMDB

Кожен запис метаболітів містить більше 90 полів даних (рисунок 6), причому половина інформації присвячена хімічним або фізико-хімічним даним, а інша – біологічним або біомедичним даним (хвороба, концентрація біологічної рідини, фермент, ген, SNP або метаболізм).

Окрім надання комплексних числових, послідовних і текстових даних, кожен запис також містить гіперпосилання на інші бази даних (рисинок 7), такі як KEGG, BioCyc, PubChem, ChEBI, PubMed, PDB, SwissProt, GenBank, OMIM

і dbSNP, реферати, цифрові зображення та аплети для перегляду молекулярних структур (рис. 3.4).

Metabolite and medical information	Protein/enzyme information
Common name	Enzyme/protein name
Description	Enzyme/protein synonyms
Synonyms/IUPAC name	Enzyme/protein sequence
Chemical structure	Protein number of residues
Chemical taxonomy	Protein molecular weight
Molecular weight (mono and ave)	Protein pI
SMILES (isomeric and canonical)	Protein gene ontology
KEGG/PubChem/OMIM/MetaGene links	Protein general function
CAS number	Protein specific function
InChi identifier	Protein pathways
Melting point	Protein reactions
Water solubility (predicted and expt)	Protein Pfam domains
State (solid, liquid, gas)	Protein signal sites
pKa or pI	Protein transmembrane regions
LogP or hydrophobicity	Protein metabolic importance
MOL/SDF/PDF text files	Protein/enzyme EC link
MOL/PDB image files	GenBank, SwissProt, PDB ID
NMR spectra (predicted, calculated)	Protein structure data
Location (cell, biofluid, tissue)	Protein cellular location
Concentration (urine, plasma, CSF)	Gene sequence
Associated disorders	GenBank ID
Abnormal concentration (urine, plasma, CSF)	Chromosome location
Metabolic pathways (KEGG, SimCell)	Chromosome locus
Metabolizing enzymes	Protein/enzyme SNPs/mutations
Metabolizing ENZYMES	Protein/enzyme references

Рисунок 3.4 – Дані, що можна знайти, перейшовши на сторінку метаболіту.

Ключовою особливістю, що відрізняє HMDB від інших метаболічних ресурсів, є його підтримка функцій пошуку і вибору баз даних вищого рівня. На додаток до вже описаних функцій перегляду та сортування даних, HMDB також пропонує утиліту пошуку хімічної структури, програма пошуку гомологів BLAST, яка підтримує запити як з однією, так і з декількома послідовностями, а також логічний текстовий пошук на основі GLIMPSE, інструмент вилучення реляційних даних, інструмент спектрального зіставлення MS і інструмент для пошуку ЯМР спектра (для ідентифікації

сполук за допомогою даних MS або ЯМР з інших метаболічних досліджень) (рис.3.5).

External Links	
DrugBank ID	DB04151
Phenol Explorer Compound ID	Not Available
FoodDB ID	FDB093588
KNApSAcK ID	Not Available
Chemspider ID	83153
KEGG Compound ID	C01152
BioCyc ID	Not Available
BiGG ID	Not Available
Wikipedia Link	Methylhistidine
METLIN ID	3741
PubChem Compound	92105
PDB ID	Not Available
ChEBI ID	50599
Food Biomarker Ontology	Not Available
VMH ID	Not Available

Рисунок 3.5 – Посилання на інші бази даних у HMDB

Інструмент пошуку схожості структури HMDB (ChemQuery) є еквівалентом BLAST для хімічних структур. Користувачі можуть робити ескіз за допомогою вільно доступного аплету ChemSketch (ACD) або вставити рядок SMILES сполуки запиту до вікна Chem-Query (рис. 3.6).

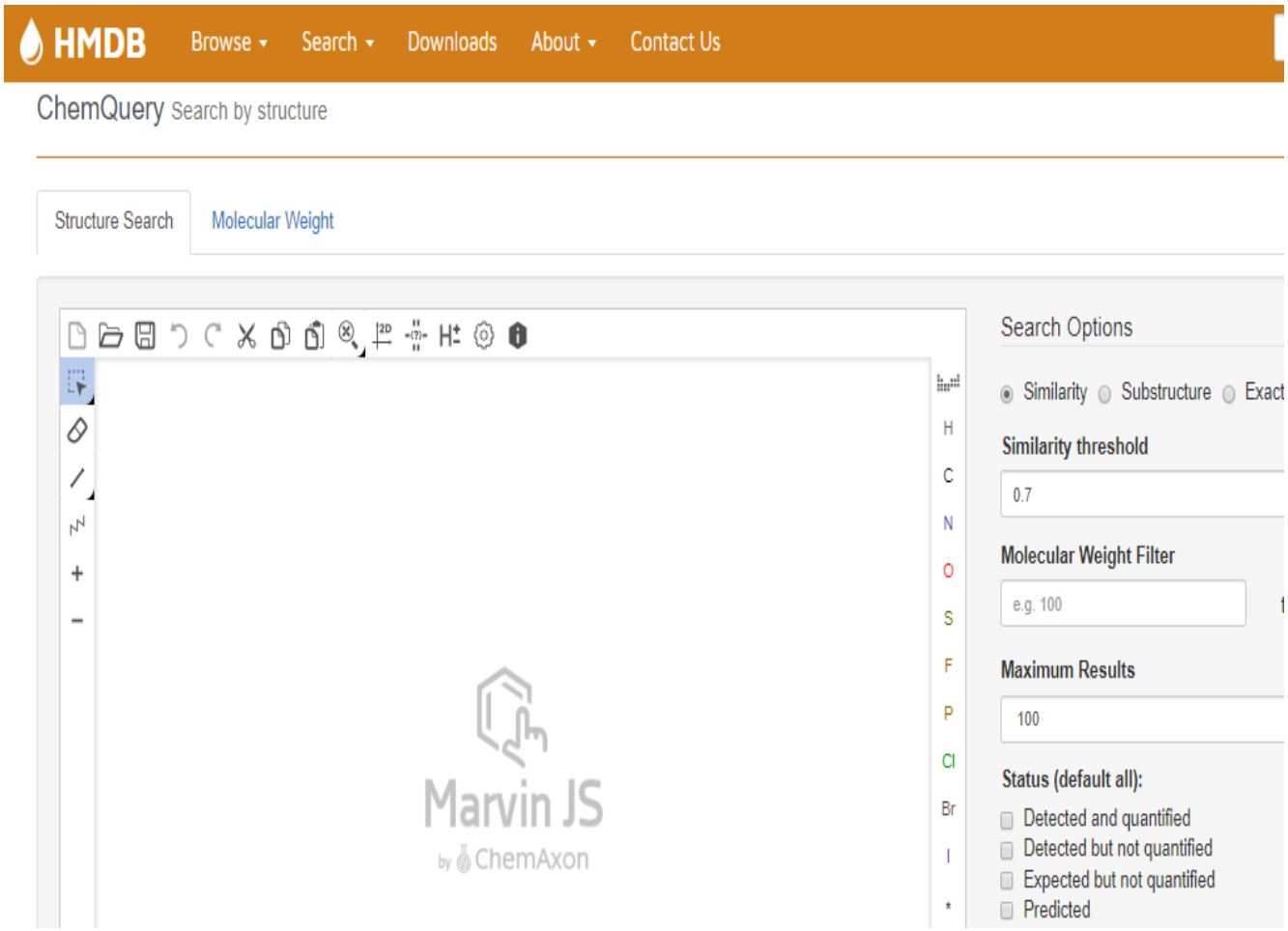


Рисунок – 3.6 Інструмент пошуку хімічних структур Chem-Query

Відправка запиту запускає інструмент пошуку схожої структури, який шукає спільні підструктури із сполуки-запиту, які відповідають базі даних метаболітів HMDB. Результати з високими співпадіннями представляються у вигляді таблиці з гіперпосиланнями на відповідні MetaboCards (які, у свою чергу, посилаються на цільовий білок). Інструмент ChemQuery дозволяє користувачам швидко визначити, чи є їхня сполука відомим метаболітом або хімічно пов'язаною з відомим метаболітом. На додаток до цих пошуків подібності структури, утиліта ChemQuery також підтримує пошук сполук на основі хімічних формул і діапазонів молекулярної маси (рис.3.7).

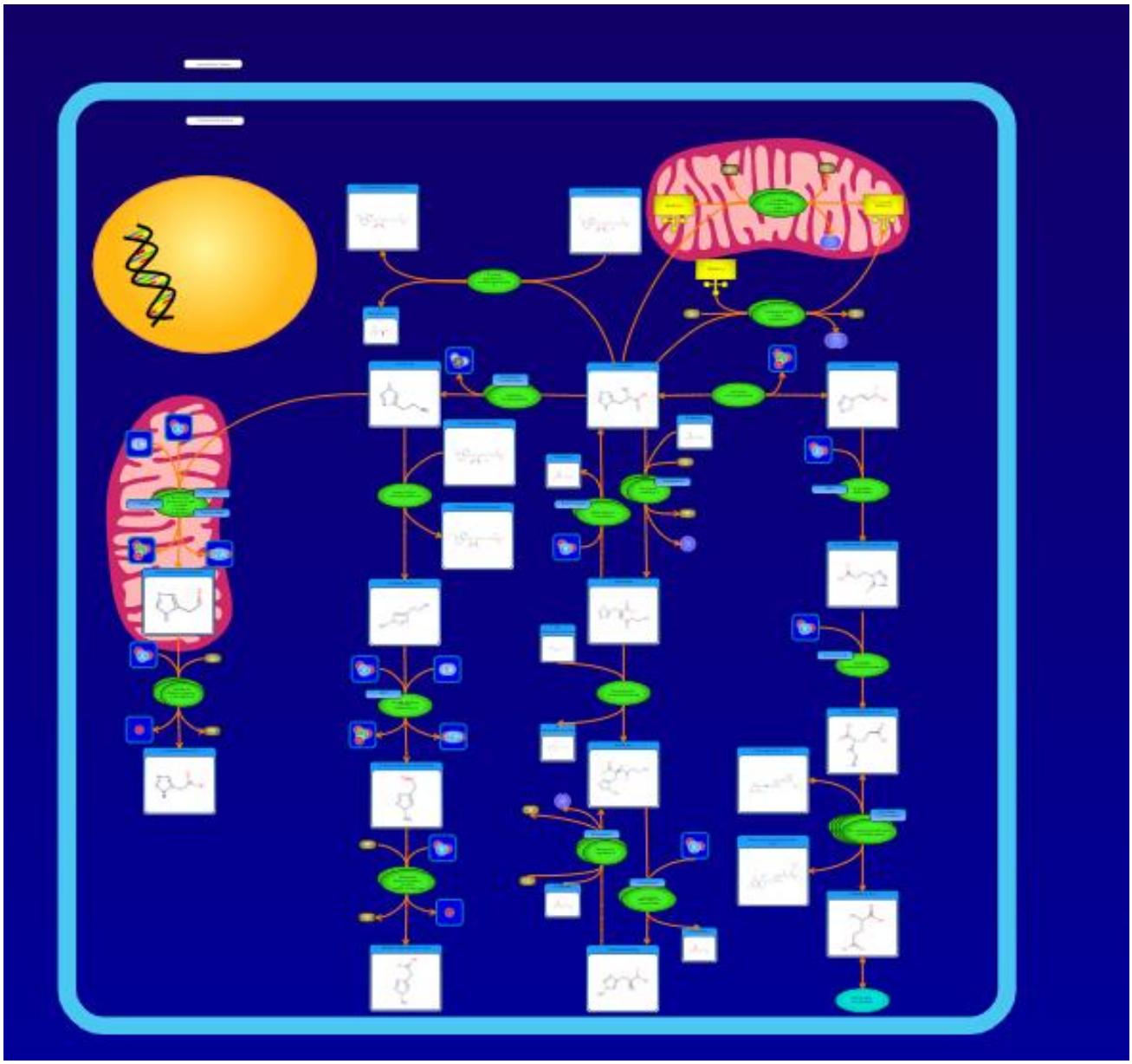


Рисунок 3.7 – Зображення метаболічних шляхів гістидину, отримане за допомогою HMDB

Програма BLAST (Sequence Search) дозволяє користувачам здійснювати пошук в HMDB через подібність послідовностей, а не за хімічною подібністю. Певну генну або білкову послідовність можна знайти на основі бази даних послідовностей HMDB метаболічно важливих ферментів і транспортерів, вставивши послідовності у форматі FASTA, в поле запиту Sequence Search і натиснувши кнопку "search" (рис. 3.8).

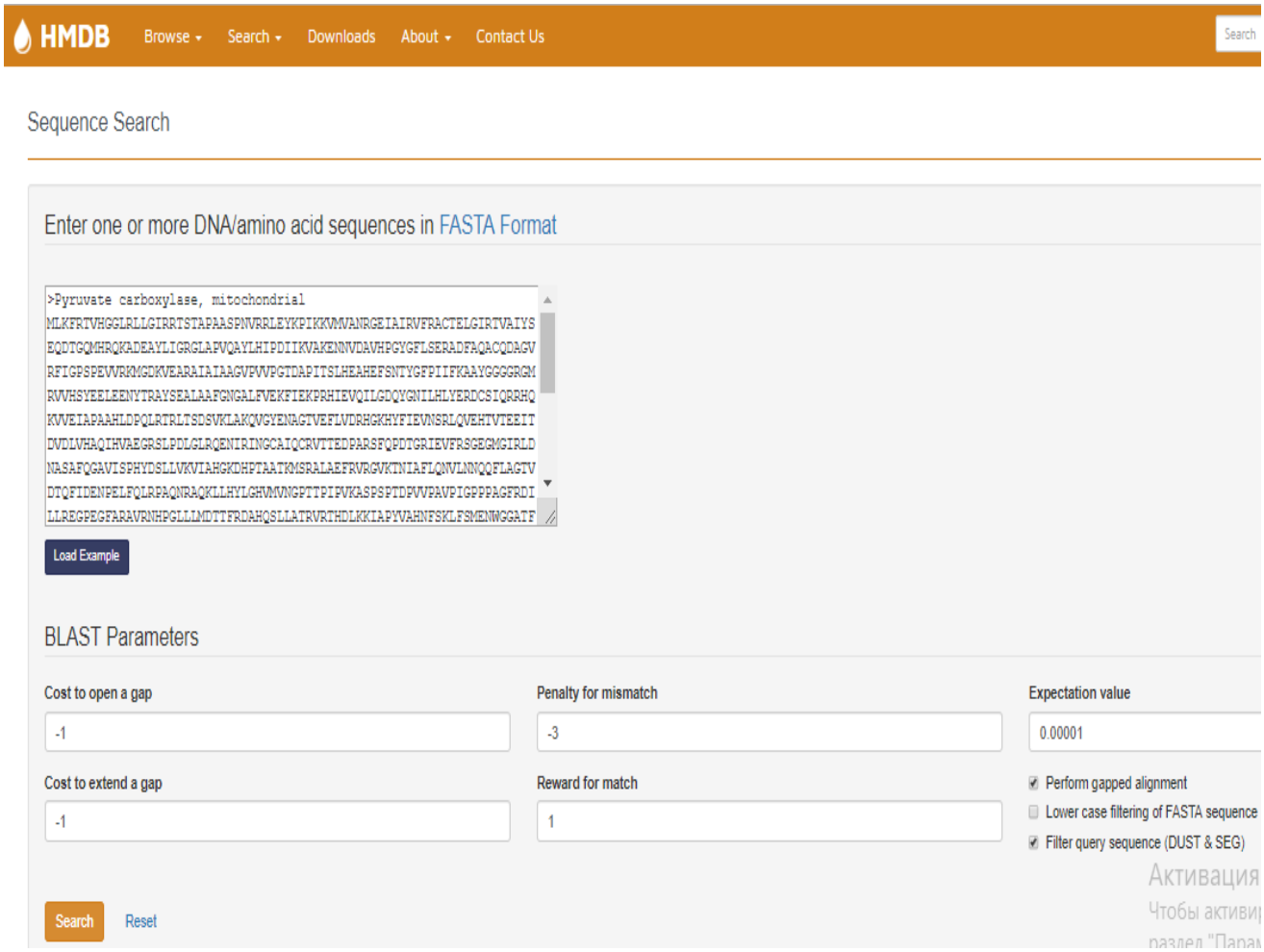


Рисунок 3.8 – Інструмент Sequence Search (BLAST) у HMDB

Значна схожість виражається через пов'язану гіперпосиланням MetaboCard, ім'я та (або) хімічну структуру метаболітів, які можуть впливати на білок, що аналізується. За допомогою взаємодії метаболіт-білок інструмент Sequence Search (BLAST) нещодавно секвенованих ссавців (шимпанзе, щура, миші, собаки, кішки тощо).

Утиліти пошуку ЯМР та MS дозволяють користувачам завантажувати спектри (для пошуку MS) або списки піків (для пошуку ЯМР) і шукати відповідні сполуки з колекції HMDB MS і ЯМР-спектрів. HMDB містить приблизно 3800 прогнозованих спектрів ЯМР для 1900 сполук. Передбачені спектри ЯМР генерували з використанням програмного забезпечення ACD/HNMR та ACD/CNMR від Advanced Chemistry Development Inc. Крім

того, HMDB містить 930 експериментально зібраних спектрів ЯМР для 400 чистих сполук. Вона також містить 1200 MS/MS (Triple-Quad) спектрів при трьох різних енергій зіткнень для майже 400 чистих сполук. Щомісяця додаються в середньому 50 нових спектрів ЯМР та MS. Спектральні пошукові засоби HMDB дозволяють ідентифікувати як чисті сполуки, так і суміші сполук з їх MS або ЯМР-спектрами через алгоритми відповідності піку, які були розроблені авторами HMDB. Алгоритм спектрального порівняння ЯМР використовує просте правило порівняння піку з попередньо визначеними допусками хімічного зсуву. Спектри запитів оцінюються за кількістю пікових збігів зі спектрами бази даних. Алгоритм MS/MS спектрального порівняння використовує концепцію пікового збігу та спектральної оцінки. Повний набір анатованих спектральних зображень (ЯМР та MS, як експериментальні, так і передбачені) активуються як "zip filesthrough" через кнопку "Download", що розташована у верхній частині меню HMDB.

Посилання «Metabolite Library» (HML) у меню HMDB «Browse» – бібліотека метаболітів людини (рис. 12). Це сховище всіх придбаних, синтезованих і ізольованих метаболітів, які були отримані командою HMP (Human Metabolome Project). Невеликі кількості окремих сполук або більших колекцій метаболітів можуть бути придбані (за певною ціною) або вільно отримані для спільних досліджень (через угоди про передачу матеріалів) через веб-сайт HML та його веб-форми замовлення. Ці сполуки можуть бути використані в якості еталонних або кількісних стандартів дослідниками метаболоміки, або колекції можуть бути використані для скринінгу лікарських засобів, скринінгу кристалів і аналізу функцій ферментів.

Можливо, найбільш важливими особливостями HMDB з точки зору медичного генетика або клінічного хіміка є її багатий вміст і широкий зв’язок із метаболічними захворюваннями, з нормальними та аномальними діапазонами концентрацій метаболітів (у багатьох різних біологічних рідинах), з даними мутації / SNP і з генами, ферментами, реакціями та шляхами, що пов'язані з багатьма захворюваннями. В даний час HMDB

містить 115 діаграм метаболічних шляхів або метаболічних карт. Хоча це число може здаватися невеликим, загальна кількість відомих шляхів людини в базі даних KEGG становить всього 190, причому 72 з них є білковими шляхами (тобто не мають метаболітів). Тим не менш, очікується, що ця кількість зросте, оскільки зростає кількість нових шляхів регулювання метаболізму генів, що визначаються за допомогою геномних досліджень, багато з яких будуть включені до HMDB.

Нещодавнім доповненням до HMDB є серія метаболічних електричних схем SimCell, моделі SimCell в SBML (Language Biology Markup Language) і SimCellsimulating майже 30 добре описаних метаболічних шляхів. SimCell – це пакет програмного забезпечення для моделювання метаболізму, який дозволяє моделювати складні метаболічні шляхи на клітинному рівні, а також створювати відео ферментативних процесів у реальному часі. Наявність цих попередньо зібраних метаболічних моделей повинно дозволити користувачам просто завантажити схему підключення SimCell і провести експерименти випробування гіпотез «*in silico*» щодо можливих причин можливого генетичного захворювання.

3.2.2 Забезпечення якості та повноти бази даних HMDB

Метаболом людини не так добре вивчений, як геном людини. Склад метаболому залежить від того, як визначається метаболіт, включаючи його поріг молекулярної маси, походження та межу концентрації. Якщо кожну малу молекулу в організмі (харчову добавку, рослинний екстракт, лікарський засіб, похідне лікарського засобу, токсин, чистячий засіб або забруднювач навколишнього середовища) слід включати з будь-якого джерела на будь-якому рівні концентрації, кількість сполук може перевищувати 100 000. Для досягнення актуальності та обґрунтованості бази даних необхідні критерії для включення в HMDB, що полягають у наступному: сполука має молекулярну масу <1500 Да, в концентраціях більше 1 мМ (або в нормальному стані або в

умовах хвороби) в одній або декількох біологічних рідинах/тканинах і має бути біологічного походження (або генеруватися клітинами людини або ендогенною мікрофлорою кишечника). Існують деякі винятки, включаючи низький вміст, щодо важливих біомедичних метаболітів (гормони, метаболіти, пов'язані з хворобами, основні поживні речовини і сигнальні молекули), деякі дуже поширені препарати (ацетамінофен, нікотин) і деякі розповсюджені харчові добавки (целобіоза, вітаміни).

Під час анотування HMDB робиться все можливе для забезпечення максимальної повноти, правильності та актуальності інформації. Метаболіти спочатку ідентифікуються за допомогою літературних оглядів (PubMed, OMIM, OMMBID, підручники), інтелектуального аналізу даних (KEGG, Metlin, BioCyc) або експериментальних методів. Якщо сполука проходить критерії включення в HMDB, інформація про метаболіт вводиться або готується одним членом команди кураторів і окремо затверджується другим членом команди кураторів. Додаткові перевірки регулярно проводяться при кожному вводі нових метаболітів старшими членами групи кураторів, включаючи двох біохіміків на рівні PhD. Для анотації також використовується декілька пакетів програмного забезпечення, включаючи інструменти аналізу тексту, калькулятори хімічних параметрів та інструменти анотації білків, спочатку розроблені для DrugBank, але змінені для HMDB. Ці інструменти збирають та відображають текст (і зображення) з декількох джерел, що дозволяє кураторам порівнювати, оцінювати, вводити та коригувати інформацію про метаболіти або ферменти чи гени. Вихідні дані та програмні засоби, що використовуються в процесі анотації, більш докладно описані у розділі "About", розташованому в рядку меню HMDB.

Всі дані вводяться в централізовану лабораторну систему управління інформацією (LIMS), що дозволяє контролювати, змінювати і автоматично переносити всі зміни до HMDB. Перевірку узгодженості (молекулярна вага збігається з хімічною формулою, стан, що відповідає точці плавлення, відсутність негативних молекулярних мас, форматування назв є

правильним і т.д.) виконують кожну ніч за допомогою автоматизованого скрипту перевірки/виправлення. Друга система відстеження тексту впроваджена для моніторингу та відображення актуальної статистики про кількість метаболітів, ферментів та інших статистичних даних HMDB. Ця інформація відображається на сторінці "Download". Білки, які діють на метаболіти або зв'язуються з метаболітами в HMDB, ідентифікуються та підтверджуються з використанням декількох джерел (PubMed, KEGG, Bio-Cyc, підручники), як і всі структури, концентрації та шляхи метаболізму (KEGG, PubChem, журнали, підручники). Докладаються всі зусилля для перевірки даних про концентрацію за допомогою незалежних експериментальних методів.

3.3 База даних DrugBank

БД DrugBank містить еквівалентну інформацію про ~ 2280 ліків і метаболітів лікарських засобів, T3DB містить інформацію про ~ 3670 загальних токсинів і забруднювачів навколишнього середовища, SMPDB містить діаграми матаболічних шляхів і шляхів хвороб людини ~ 25000, в той час як FooDB містить еквівалентну інформацію про ~ 28000 харчових компонентів і харчових добавок.

DrugBank (рис.3.9) і FooDB (рис. 3.10) намагаються охопити такі ділянки метаболізму як екзогенні малі молекули, які дещо відрізняються від того, що можна вважати ендогенним або «істотним» для життя. У багатьох випадках (особливо в клінічній хімії) ці екзогенні хімікати або ксенобіотики можуть вважатися забруднюючими речовинами. В інших випадках (особливо в галузі харчування та досліджень лікарських засобів) вони являють собою інтерес. Очевидно, що одним з основних класів екзогенних або ксенобіотичних метаболітів, що зустрічаються майже у всіх людей, є лікарські засоби.

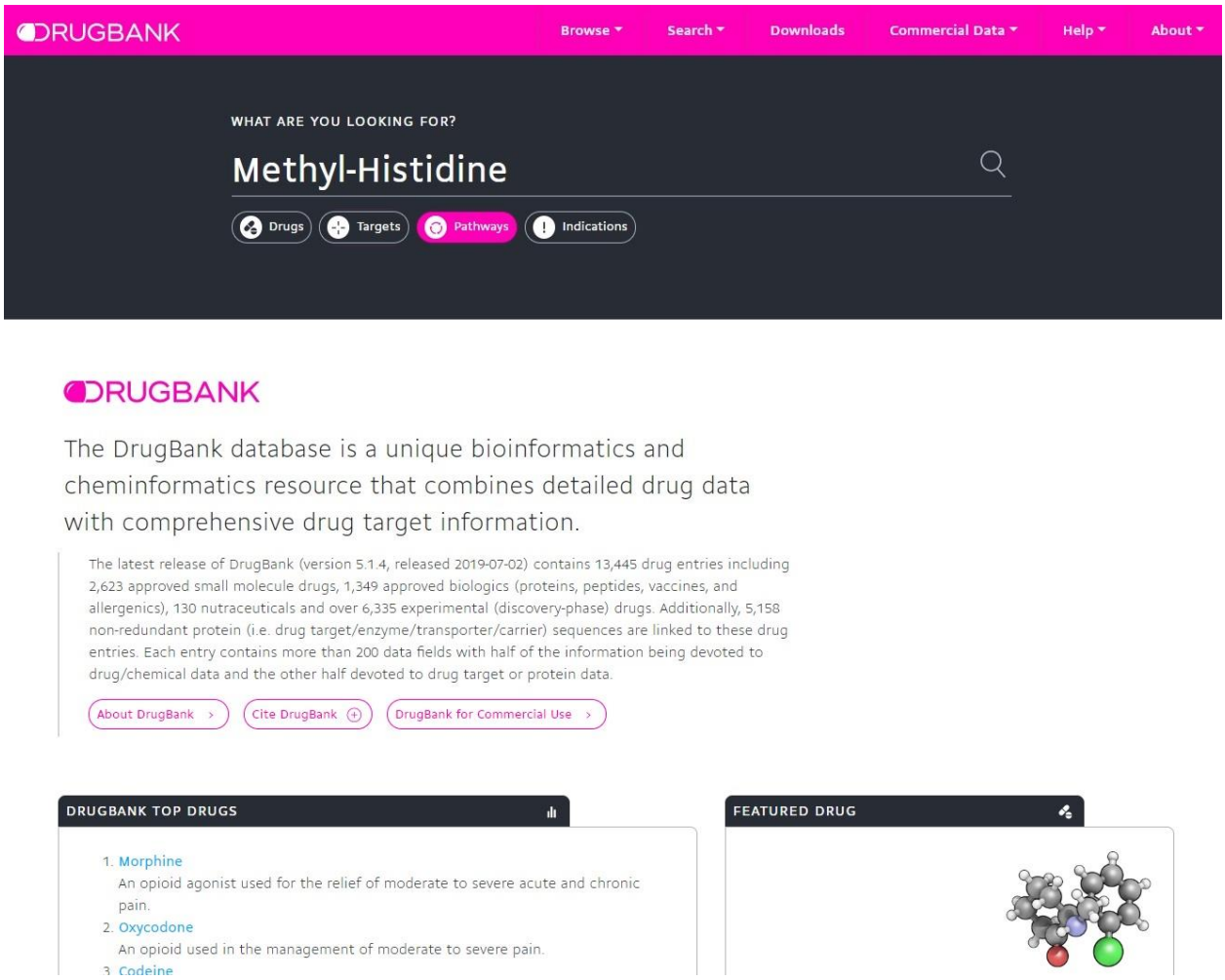


Рисунок 3.9 – Стартова сторінка DrugBank

Споживання лікарських засобів, як законних, так і незаконних, особливо поширене в розвинених країнах світу. Приблизно 80% дорослих в США використовують принаймні один безрецептурний або рецептурний лікарський засіб на тиждень, причому 25% літніх людей приймають більше п'яти рецептурних ліків на тиждень.

Споживання алкоголю, тютюну та вітамінів або біологічно-активних добавок майже однаково широко розповсюджене. Це означає, що ліки повинні розглядатися, як важлива частина метаболізму людини. Для отримання цієї інформації і було розроблено окрему базу даних лікарських засобів DrugBank. DrugBank містить детальні фізико-хімічні, фармакологічні та біохімічні дані про більш ніж 4300 препаратів, у тому числі більше 1100 лікарських засобів,

затверджених FDA, і більше, ніж 3200 експериментальних і заборонених лікарських засобів.

DrugBank структурований аналогічно HMDB, з аналогічними переглядами даних і функціями запитів. Кожен запис лікарського засобу містить більше 80 різних текстових полів, зображень або гіперпосилань, включаючи описи лікарських засобів, способи дії, дозування, кінетику, ферменти та мішені. Найбільш важливим є те, що більш ніж 16000 послідовностей лікарського препарату (тобто білків) пов'язані із записами лікарських засобів DrugBank, багато з яких мають детальні дані SNP поряд з 3D-структурами або моделями 3D гомології, пов'язаними з ними.

3.4 База даних FooDB

З метою охоплення метаболізму поживних речовин було розроблено окрему базу даних FooDB (рисунок 3.10).

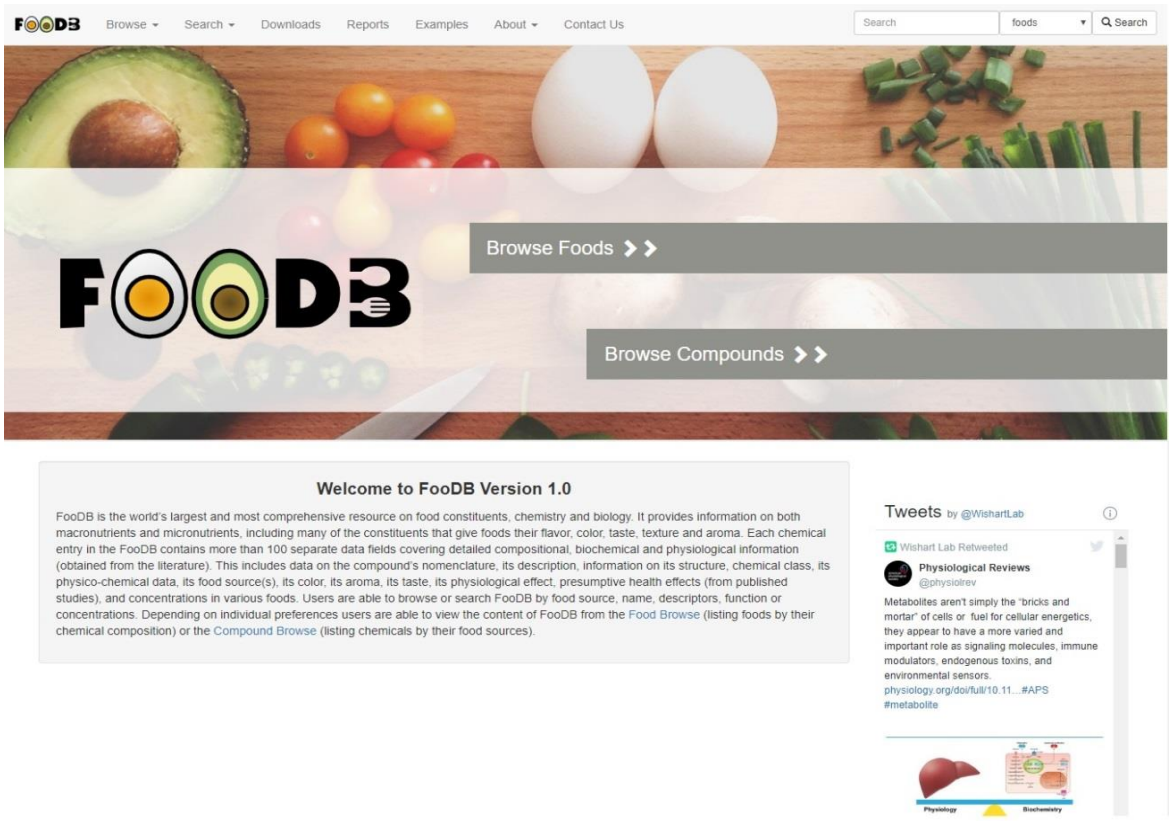


Рисунок .10 – Стартова сторінка FooDB

Лікарські засоби та метаболіти лікарських засобів не є єдиними ксенобіотичними речовинами, що зустрічаються у людей. Будучи всеїдними, люди споживають широкий асортимент рослинних, тваринних і мікробних продуктів, включаючи різноманіття (більше 3000) різних консервантів, спецій, ароматизаторів, харчових добавок і синтетичних харчових добавок. Багато з цих синтетичних добавок і екзотичних рослинних продуктів включаються до складу наших клітин або виводяться з організму як відходи в різних біологічних рідинах. Іншими словами, вживаючи їжу, люди, по суті, роблять метаболічну частину іншого організму нашою власною. Поширеність цих «чужорідних» харчових продуктів в клітинах, тканинах і біологічних рідинах може призвести до заплутаних результатів у багатьох клінічних дослідженнях і експериментах. Однак відомо, що деякі з цих чужорідних сполук (наприклад, ресвератрол і сульфорафан) та їх метаболіти мають важливі терапевтичні ефекти, тому їх визначення являє собою значний інтерес для спеціалістів-дієтологів.

Вона побудована з використанням тієї ж самої технології та анотацією з тими ж інструментами для обробки даних, як DrugBank і HMDB, FooDB наразі містить хімічні та біохімічні дані про велику кількість харчових добавок та харчових компонентів. Краща версія цієї бази даних тепер доступна на веб-сайтах HMDB та DrugBank.

Резвератрол – фітоалексин, що в природних умовах виробляється кількома рослинами за умовами інфекції бактеріями або грибами.

Фітоалексини – антибактеріальні та протигрибкові речовини, що виробляються рослинами у відповідь на інфекцію патогенів. Резвератрол також може бути бути отриманий за допомогою хімічного синтезу та продається як біологічно активна добавка. Резвератрол має багато корисних ефектів на здоров'я людини та тварин, повідомлені його дія проти раку, вірусних інфекцій, старіння, протизапальна дія, та ефект збільшення тривалості життя, хоча багато з цих результатів отримані на тваринах (наприклад, щурах), і не були підтверджені на людині. Резвератрол

знаходиться в шкірці червоного винограду і червоному вині, але засновуючись на дослідженнях тварин, в занадто малих кількостях для пояснення «французького парадоксу», тобто відносно низької частоти серцевих хвороб у жителів півдня Франції, незважаючи на високе споживання насичених жирів.

Сульфорафан – протиракова і протимікробна речовина, що міститься в деяких рослинах родини капустяних та пов'язаних видах, таких як брюсельська
капуста, броколі, капуста, цвітнакапуста, бокчой, кольрабі, турнепс, редиска. ФФермент мірозиназа трансформує *глюкорафанін* (глюкозінолат) в сульфорафан при механічному пошкодженні рослини, наприклад, пережовуванні. Молоді пагони броколі та цвітної капусти особливо багаті глюкорафанін.

Сульфорафан збільшує виробництво ензимів, які виводять токсини з організму. Чим молодшою є брокколі, тим вище в ній вміст сульфорафану.

3.5 Бази даних сполук

Бази даних сполук не містять інформації про метаболічні шляхи, а скоріше зосереджені на наданні детальних номенклатурних, структурних або фізико-хімічних даних щодо обмежених класів сполук, таких як ліпіди, вуглеводи, лікарські засоби, токсини та інші хімічні речовини, залучені до біологічних процесів. Ці дещо спеціалізовані бази даних часто містять метаболіти або ксенобіотики, які не зустрічаються у більшості баз даних метаболічних шляхів.

1. ChEBI (Chemical Entities of Biological Interest) – вільнодоступний словник молекулярних об’єктів, орієнтований на невеликі хімічні сполуки. Це або природні метаболіти або синтетичні продукти, що використовуються для втручання в процеси живих організмів. Містить інформацію про структуру та номенклатуру, а також гіперпосилання на інші відомі бази даних.

2. PubChem – вільнодоступна база даних з хімічними структурами малих органічних молекул та інформацією про їх біологічну активність. Містить структуру, номенклатуру та розраховані фізико-хімічні данні.

3. ChemSpider – база даних органічних молекул. Має розширені можливості пошуку, а для більшості сполук розраховані значення фізико-хімічних властивостей.

4. KEGG Glycan – містить набір експериментально визначених структур гліканів великої кількості еукаріотичних та прокаріотичних організмів.

5. IIMDB (*In Vivo/In Silico* Metabolites Database) – база даних, що містить як відомі, так і вирахувані сполуки – метаболіти ссавців, лікарські засоби, вторинні метаболіти рослин та гліцерофосфоліпіди.

6. DrugBank – змішаний ресурс біоінформатики та хімічної інформатики, який поєднує детальну інформацію про лікарські засоби (хімічну, фармакологічну та фармацевтичну) та всебічну інформацію про цільовий лікарський засіб (структура та шляхи розповсюдження).

3.6 Бази даних спектрів

Бази даних спектрів містять еталонні спектри ЯМР (ядерно-магнітний резонанс), GC-MS (газова хроматографія-мас-спектроскопія) та/або LC-MS (рідинна хроматографія-мас-спектроскопія) для великої кількості малих молекул разом із програмним забезпеченням для ідентифікації цих сполук за допомогою спектральної відповідності.

1. BMRB (BioMagResBank) – центральне сховище експериментальних ЯМР даних, насамперед для макромолекул. Також містить розділ для даних про метаболіти.

2. MMCD (Madison Metabolomics Consortium Database) – база даних малих молекул, що зібрана із електронних баз даних та наукової літератури. Містить інформацію про хімічну формулу, назви, синоніми, структуру,

фізичні та хімічні властивості, дані ЯМР та MS для чистих сполук, наявність метаболіту у різних біологічних видів.

3. MassBank – мас-спектральна база даних експериментально отриманих спектрів MS метаболітів.

4. Golm Metabolome Database – забезпечує відкритий доступ до бібліотек GC/MS.

5. Metlin – сховище мас-спектральних даних метаболітів. Всі метаболіти є нейтральними або вільними кислотами. Містить дані MS/MS, LC/MS та FTMS, які можна шукати за діапазоном мас, біологічним джерелом та захворюванням.

Запитання до розділу 3

1. Які події сприяли появі нових наукових напрямків таких як геноміка, протеоміка, метаболоміка, фармакогеноміка?
2. Дати визначення генній терапії.
3. Назвіть принципову відмінність генної терапії від традиційних способів лікування.
4. Назвіть основні доступі проекти в сфері персоналізованої медицини.
5. Які БД спеціалізованих біоінформатичних ресурсів сприяють розвитку послуг персоналізованої медицини?
6. Дати визначення метаболоміки.
7. Назвіть основні типи баз даних, що використовуються в дослідженнях метаболоміки.
8. Назвіть найбільш відомі БД метаболічних шляхів.
9. Назвіть найбільш відомі БД сполук.
10. Які БД містять еталонні спектри ЯМР, GC-MS (газова хроматографія-мас-спектроскопія) та/або LC-MS (рідинна хроматографія-мас-спектроскопія)?
11. Яку інформацію містять БД захворювань?

12. Назвіть БД, яка описує генетику, метаболізм, діагностику та лікування порушень обміну речовин.
13. Яка БД містить детальну інформацію про метаболіти малих молекул, виявлених (та підтверджених експериментально) в організмі людини?
14. Мета створення БД FooDB?
15. Назвіть БД, яка містить еквівалентну інформацію про ~ 2280 ліків і метаболітів лікарських засобів.

Література до розділу 3

1. BioMagResBank / [E. L. Ulrich, H. Akutsu, J. F. Doreleijers та ін.]. // Nucleic Acids Research. – 2008. – №36. – С. 408–408.
2. Chadwick, R., & O'connor, A. (2013). Epigenetics and personalized medicine: prospects and ethical issues. Personalized Medicine, 10(5), 463-471
3. ChEBI: a database and ontology for chemical entities of biological interest / [K. Degtyarenko, P. de Matos, M. Ennis та ін.]. // Nucleic Acids Research. – 2008. – №36.
4. Database resources of the National Center for Biotechnology Information / [D. L. Wheeler, T. Barrett, D. A. Benson та ін.]. // Nucleic Acids Research. – 2007. – №35. – С. 5–12.
5. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Detailed description of DrugBank and its potential applications / Wishart DS, Knox C, Guo A та ін.]. // Nucleic Acids Res. – 2006. – №34. – С. 668–672.
6. Farina A, Ferranti C, Marra C An improved synthesis of resveratrol. Nat. Prod. Res. 20 (3), (2006).
7. FooDB food component database. URL: http://hmbd.med.ualberta.ca/foodb
8. Frauendienst-Egger G. Metagene – knowledge base for inborn errors of metabolism (3.0) / G. 15. Frauendienst-Egger, F. K. Trefz. // Indian J Pharmacol. – 1999. – №31. – С. 321.

9. From genomics to chemical genomics: new developments in KEGG / [M. Kanehisa, S. Goto, M. Hattori та ін.]. // Nucleic Acids Research. – 2006. – №34. – С. 354–357.
10. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs / Altschul S.F., Madden T.L, Schaffer A.A. та ін.]. // Nucleic Acids Res. – 1997. – №25. – С. 3389–3402.
11. German J. B. Metabolomics: building on a century of biochemistry to guide human health / J. B. German, B. D. Hammock, S. M. Watkins. // Metabolomics. – 2005. – №1. – С. 3-9.
12. Ginsburg, G. S., & Willard, H. F. (2009). Genomic and personalized medicine: foundations and applications. Translational research, 154(6), 277-287
13. GMD DB: the Golm Metabolome Database. / [J. Kopka, N. Schauer, S. Krueger та ін.]. // Bioinformatics. – 2005. – №21. – С. 1635–1638.
14. Heidecker, B., & Hare, J. M. (2007). The use of transcriptomic biomarkers for personalized medicine. Heart failure reviews, 12(1), 1-11
15. HMDB 4.0: the human metabolome database for 2018 / [W. S. David, Y. D. Feunang, A. Marcu та ін.]. // Nucleic Acids Res. – 2018. – №46. – С. 608–617.
16. HMDB: the Human Metabolome Database / Wishart David S., Tzur Dan, Knox Craig та ін.]. // Nucleic Acids Res. – 2007. – №35. – С. 521–526. http://www.metabolomicssociety.com/resources/metabolomics-databases.
17. Human Matabolome Database URL: http://www.hmdb.ca/
18. Manber U. WebGlimpse—Combining Browsing and Searching, Usenix / Manber U., Smith M., Gopal B.. // Annual Technical Conference, Anaheim, CA. – 1997. – С. 195–206.
19. Metabolite identification via the Madison Metabolomics Consortium Database / [Q. Cui, I. A. Lewis, A. D. Hegeman та ін.]. // Nature Biotechnology. – 2008. – №26. – С. 162–164.
20. METLIN: a metabolite mass spectral database / [C. A. Smith, G. O'Maille, E. J. Want та ін.]. // Therapeutic Drug Monitoring. – 2005. – №27. – С. 747–751.

21. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders / [A. Hamosh, A. F. Scott, J. S. Amberger та ін.]. // Nucleic Acids Research. – 2005. – №33. – С. 514–517.
22. Quackenbush J. Extracting biology from high-dimensional biological data / Quackenbush. // Journal of Experimental Biology. – 2007. – №210. – С. 1507–1517.
23. Querying and computing with BioCyc databases / [M. Krummenacker, S. Paley, L. Mueller та ін.]. // Bioinformatics. – 2005. – №21. – С. 3454–3455.
24. Reactome: a knowledgebase of biological pathways. / [G. Joshi-Tope, M. Gillespie, I. Vastrik та ін.]. // Nucleic Acids Research. – 2005. – №33. – С. 4.
25. Rivenbark, A. G., O'Connor, S. M., & Coleman, W. B. (2013). Molecular and cellular heterogeneity in breast cancer: challenges for personalized medicine. The American journal of pathology, 183(4), 1113-1124
26. The Universal Protein Resource (UniProt) / Bairoch A., Apweiler R., Wu C.H. та ін.]. // Nucleic Acids Res. – 2005. – №33. – С. 154–159.
27. Van Der Greef, J., Hankemeier, T., & McBurney, R. N. (2006). Metabolomics-based systems biology and personalized medicine: moving towards n= 1 clinical trials?. Pharmacogenomics, 7(7), 1087-1094
28. Virgin, H. W., & Todd, J. A. (2011). Metagenomics and personalized medicine. Cell, 147(1), 44-56
29. Weininger D. SMILES 1. Introduction and Encoding Rules / Weininger D. // Journal of chemical information and computer sciences. – 1988. – №28. – С.31–38.
30. Weston, A. D., & Hood, L. (2004). Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. Journal of proteome research, 3(2), 179-196
31. Wishart D. Human Metabolome Database: Completing the "Human Parts List" / David S. Wishart. – 2007. – №8. – С. 683–686.

РОЗДІЛ 4. БД МЕДИЧНОГО СПРЯМУВАННЯ

4.1 БД захворювань і фізіології

Бази даних захворювань містять описи причин, клінічні симптоми, діагностичні показники або генетичні мутації, пов’язані з багатьма порушеннями обміну речовин.

1. OMIM (Online Mendelian Inheritance in Man) – всебічний збірник генів людини та генетичних фенотипів. Зосереджується на взаємозв’язку генотипу та фенотипу. Містить захворювання із супутніми даними про послідовність генів, фенотипи захворювань та відомі їх генетичні причини.
2. METAGENE – база даних з інформацією про вроджені помилки метаболізму, що надає інформацію про захворювання, генетичну причину, лікування та характерні концентрації метаболіту або клінічні тести, які можуть бути використані для діагностики або моніторингу стану, а також є дані про генетичні захворювання.
3. OMMBID (On-Line Metabolic and Molecular Basis to Inherited Disease) – інтернет-доступна енциклопедія, що описує генетику, метаболізм, діагностику та лікування сотень порушень обміну речовин.

4.2 БД однонуклеотидних поліморфізмів (SNP)

4.2.1 Загальна характеристика SNP

Однонуклеотидний поліморфізм (англ. Single nucleotide polymorphism, SNP, вимовляється як «СНП» або «СНіП») – відмінності послідовності ДНК розміром в один нуклеотид (A, T, G або C) в геномі (або в іншій послідовності, що порівнюється) представників одного виду або між гомологічними ділянками гомологічних хромосом, де кожна варіація присутня на рівні понад 1% у популяції. Поліморфізм – це одночасне існування в популяції кількох

алельних варіантів будь-якого гена. Одним з найбільш представлених в геномі людини поліморфізмів є саме однонуклеотидний поліморфізм. Близько 90% варіацій генома обмежуються SNP, які, як було доведено, мають велике значення для медичної діагностики і розробки фармацевтичних продуктів. Однак, досить часто, всупереч формальному визначенню, всі невеликі зміни в геномних послідовностях, виявлені в ході SNP скринінгу, розміщуються в одних і тих же базах даних. В одному списку з SNPs виявляються невеликі інсерції/делеції (indels) і зміни декількох нуклеотидів.

SNP – це найбільш поширені генетичні відмінності між людьми, вважається багатообіцяючим шляхом до персоналізованої медицини. За даними досліджень функціональної геноміки однонуклеотидні заміни в смислових ділянках гена в більшості випадків впливають на експресію, тим самим, змінюючи такі характеристики білка, як третинна структура, стабільність зв'язування з субстратом і проміжними метаболітами, посттрансляційні модифікації. При цьому функціональні характеристики білків можуть сильно змінюватися від практично нейтрального ефекту генетичного поліморфізму до повного порушення функції відповідного білкового продукту. Зміна рівня експресії гена може слугувати причиною різних захворювань. Одним з можливих джерел зміни експресії може бути наявність однонуклеотидного поліморфізму в регуляторній області, коли при певному алельному стані відбувається порушення сайту зв'язування транскрипційного фактора. Окрім цього SNP в кодуючих районах генів і сайтах сплайсингу призводять до зміни структури білкового продукту гена, стабільності зв'язування з субстратом і проміжними метаболітами, посттрансляційних модифікацій, реакцій у відповідь на патогени, прийом ліків, вакцин і та ін. Величезне значення SNPs в біомедичних дослідженнях полягає в тому, що їх використовують для порівняння ділянок геному між досліджуваними групами (наприклад, одна група - люди з певним захворюванням, а друга - без нього. При цьому функціональні характеристики білків можуть сильно змінюватися від практично нейтрального ефекту

генетичного поліморфізму до повного порушення функції відповідного білкового продукту. Такі SNP широко вивчаються, оскільки вони часто виявляються асоційованими з різними клінічно важливими ознаками, пояснюють характер проходження різних захворювань, вони також можуть допомогти ідентифікувати множинні гени, повязані з складними хворобами, такими як рак, діабет, тощо.

На даний час існує щонайменше 5 баз даних, що використовуються для зберігання інформації та дослідження виявлених однонуклеотидних поліморфізмів:

DbSNP - база даних SNP, вільний архів, що містить дані по спадковій мінливості різних видів, розроблений і підтримуваний NCBI (National Center for Biotechnology Information, Національний центр біотехнологічної інформації США). Хоча така назва бази даних має на увазі, що там зібрано тільки один клас поліморфізмів, саме SNPs, а насправді вона містить велику кількість інформації і про інші молекулярні зміни в амінокислотних послідовностях. dbSNP була створена у вересні 1998 року на додаток до бази даних GenBank. У 2010 році dbSNP містила більше 184 мільйонів послідовностей, представляючи більше 64 мільйонів різних варіантів для 55 організмів, включаючи *Homo sapiens, Mus musculus, Oryza sativa* і безліч інших.

SNPedia - біоінформатичний вікі-сайт, який служить як база даних SNP. Кожна стаття про SNPs надає короткий опис, посилання на наукові статті, і, крім того інформацію з мікрочіпа про однонуклеотидний поліморфізм даного типу. SNPedia допомагає в інтерпретації результатів власної генетичної інформації за допомогою таких програм як Promethease, 23andMe, Navigenics, deCODEme або Knome. SNPedia була створена і підтримується генетиком Грегом Ленноном і програмістом Майком Каріазо. До 14 вересня 2017 року у базі даних містилося 107 125 однонуклеотидних поліморфізмів.

GWAS (*genome-wide association studies, GWA study*, повногеномний пошук асоціацій) – база даних, яка характеризує короткий зміст об'єднаних

даних в одному або декількох повногеномних дослідженнях. GWAS використовує потужні графічні і текстові методи представлення даних для представлення і візуалізації багатьох однонуклеотидних поліморфізмів.

HapMap – база даних, метою якої є розвиток карти гаплотипів людського генома, яка описує загальні патерни генетичної мінливості у людей. HapMap - основний ресурс для виявлення генетичної мінливості, врахування факторів навколишнього середовища на здоров'я людини. Вся інформація, що надається знаходиться у вільному доступі. Цей проект - результат співпраці різних груп вчених з Канади, Китаю, Японії, Нігерії, Великобританії і США, остаточна версія якого побачила світ навесні 2009 року.

Kaviar - це база даних, яка включає в себе близько 160 мільйонів SNV, доповнених даними, що відносяться до коротких вставок або замін. Ця платформа покликана допомогти дослідникам у виборі SNV. Людські однонуклеотидні варіанти (SNV), зібрані з більш ніж 30 різних джерел. Пошук може бути виконаний з використанням інтегрованого інструменту, результати якого можуть бути додатково використані наступними інструментами, такими як MAGMA.

Аналіз SNP використовується у всіх областях наук про життя, включаючи молекулярну діагностику, сільське господарство, тестування продуктів харчування, тестування ідентичності, ідентифікацію патогенних мікроорганізмів, виявлення і розробку ліків та вважається багатообіцяючим шляхом до персоналізованої медицини.

4.2.2 Номенклатура та значення частоти виявлення SNPs

Єдиної номенклатури для SNPs немає: часто існують кілька різних варіантів назви для одного конкретно обраного SNP, до якоїсь згоди в цьому питанні прийти поки не вдається. По мірі того як проводиться все більше досліджень по SNP, нестандартні номенклатури можуть створювати потенційні проблеми. Найбільш серйозна проблема полягає в тому, що

дослідники не можуть виконувати перехресні посилання між різними базами даних SNP. Це приводить до збільшення ресурсів і часу, необхідних для відстеження SNP.

Один з підходів – це використання універсальної автоматизоваї системи ідентифікації SNP UASIS. UASIS – це веб-сервер для стандартизації і трансляції номенклатури SNP на рівні ДНК. Три утиліти доступні. Це UASIS Aligner, Universal SNP Name Generator і SNP Name Mapper. UASIS відображає SNP з різних баз даних, включаючи dbSNP, GWAS, HapMap і JSNP і т.д., в єдине уявлення, ефективно використовуючи запропоновану універсальну номенклатуру і сучасні алгоритми вирівнювання.

UASIS знаходиться у вільному доступі за адресою http://www.uasis.tk. UASIS є корисною платформою для перехресних посилань і відстеження SNP. Надаючи інформативну, унікальну та однозначну номенклатуру, в якій використовується унікальна позиція SNP, можливо вирішити неоднозначність номенклатур SNP, що застосовуються в даний час. Ця універсальна номенклатура є хорошим доповненням до загальноприйнятих позначень SNP, таким як rs # і HGVS.

Таблиця 4.1 представляє численні альтернативні способи позначення SNP в основних базах даних. При цьому приватні бази даних продовжують використовувати нетрадиційні номенклатури, як показано в таблиці 4.2.

Таблиця 4.1 – Альтернативні імена SNP Alternative names of a SNP

Database	**SNP names**
dbSNP	rs3737965
	ss4923964, ss69366921
HGVBaseG2P	HGVM2256489
HGVS	NM_001286.2:c.87+45G>A, NM_021735.2:c.87+45G>A
	NM_021736.2:c.87+45G>A, NM_021737.2:c.87+45G>A
	NT_021937.19:g.7871183G>A **«<Номер доступа>. <Номер версии> (<символ гена>): <тип последовательности>. <Мутация>».**
JSNP	IMS-JST083663

PharmGKB	rs3737965@chrl: 11789038
HapMap	rs3737965

Таблиця 4.2 – Нетрадиційні імена SNP Non-conventional names of a SNP

dbSNP	**rs28942082**
Genome-browser-like syntax	Chr19:11,087,877-11,087,877 G/T
	Chr19:11087877 G/T
Others	geneA,11,EXON,108,T,hetero
	gene Asynonym,11,108,exon,GT
	proteinB, Gly564Val; proteinB, Bly544Val
	014 FH NAPLES

SNP зустрічаються по всьому геному і в геномі людини зустрічаються з частотою один SNP на кожну 1000 пар основ, при середній відстані між маркерами 30тпн. У самому визначенні SNPs закладена орієнтація на їх використання в якості генетичних маркерів. Жоден інший тип геномних відмінностей не здатний забезпечити таку щільність картування, оскільки на кожен відомий або передбачуваний ген доводиться в середньому по два маркера.

Однонуклеотидний поліморфізм зустрічається в межах кодуючих послідовностей генів, в некодуючих ділянках або в ділянках між генами. SNPs, що зустрічаються в кодуючих ділянках, можуть не змінювати амінокислотну послідовність білка через виродженість генетичного коду.

В принципі, можливе існування двох/бі-, три- і чотирьох-алельних поліморфізмів. Однак на практиці надзвичайно рідкісні навіть трьохалельні SNP (менше 0.1% всіх SNP людини). Біалельні SNP можуть бути чотирьох різних типів: один вид транзицій C <=> T (<=> A) і три типи трансверсії: C <=> A (G <=> T), C <=> G (G <=> C) і T <=> A (A <=> T). Транзиції складають дві третини SNPs людини. Можливо це пов'язано з походженням C <=> T (G

<=> А) замін в реакції деамінування 5-метилцитозину.

Однонуклеотидні поліморфізми кодуючих ділянок бувають двох типів: синонімічні і несинонімічні.

Синонімічні SNPs залишають амінокислотну послідовність білка без зміни, тоді як несинонімічні SNPs змінюють її. Несинонімічні SNPs можна розділити на missense і nonsense. Однонуклеотидний поліморфізм, що зустрічається в некодуючих ділянках гена, можливо, впливає на генетичний сплайсинг, деградацію мРНК, зв'язування транскрипційних факторів.

4.2.3 Маркери SNPs

Практичний інтерес до SNPs з н а ч но зріс в ході реалізації проектів по визначенню повних нуклеотидних послідовностей ряду модельних організмів. У самому визначенні SNPs закладена орієнтація на їх використання в якості генетичних маркерів (обмеження по частоті протиставляє їх рідкісним мутацій). SNP використовується у вивченні генетичного різноманіття як альтернатива мікросателітам. Доступним є ряд технологій по виявленню і типування SNP-маркерів.

Будучи діаллельнимі маркерами, SNP мають суттєво менший інформаційний зміст, і для отримання того ж рівня інформації, який можна отримати при використанні стандартної панелі з 30 мікросателітних локусів, необхідно використовувати більшу їх кількість. Однак постійно розвиваються молекулярні технології, збільшується автоматизація і зменшується вартість типування SNP.

З такою перспективою виконуються великомасштабні проекти не тільки стосовно людини, а й по низці видів домашніх тварин для ідентифікації мільйонів і підтвердження декількох тисяч SNP, для виявлення блоків гаплотипов в геномі. Однак ефективність SNP у вивченні різноманітності у видів тварин до цих пір залишається недостатньо дослідженою.

Так само як інформація про послідовності, SNP дозволяють безпосередньо порівнювати і об'єднувати результати аналізу різних експериментів. В майбутньому SNP, мабуть, будуть привабливими маркерами для вивчення генетичної різноманітності, оскільки їх легко використовувати в оцінці і функціональної, і нейтральної мінливості.

Маркерні технології еволюціонують і, схоже, що мікросателіти послідовно заміщуються на SNP. Ці маркери дуже перспективні, оскільки число їх в геномі велике, і вони придатні для автоматизації аналізу та генотипування.

4.2.4 БД GWAS

БД GWAS спрямована на біомедичні дослідження, пов'язані з дослідженням асоціацій між геномними варіантами і фенотиповими ознаками. Часто під GWAS мають на увазі тільки пошук зв'язків між однонуклеотидним поліморфізмом і захворюваннями людини, проте термін використовується у відношенні і до інших організмів.

Однонуклеотидний поліморфізм (поряд з поліморфізмом довжин рестрикційних фрагментів) широко використовують в якості молекулярно-генетичних міток (маркерів) в повногеномному пошуку асоціацій молекулярно-генетичної систематики на основі дивергенції (розбіжності) гомологічних ділянок ДНК в філогенезі. У даній області найбільш часто використовуються спейсери генів рибосомальної РНК. З огляду на те, що мутації в даних спейсерів не позначаються на структурі кінцевих продуктів гена (теоретично вони не впливають на життєздатність), в першому наближенні постулюється пряма залежність між ступенем поліморфізму і філогенетичною відстанню між організмами.

Основна мета GWAS полягає в ідентифікації генетичних факторів ризику, щоб дати обґрунтований прогноз про схильність до захворювання, а також у виявленні біологічних основ схильності до хвороби для розробки

нових стратегій профілактики і лікування.

У дослідженнях такого типу зазвичай порівнюють геноми групи хворих людей, що мають різні фенотипи, з геномами контрольної групи, що включає в себе аналогічних за віком, статтю та іншими ознаками здорових людей. За допомогою GWAS можна порівнювати не тільки геноми пацієнтів, а й здорових людей, що мають різні прояви одної фенотипової ознаки. Матеріалом для дослідження є зразки ДНК генома кожного учасника дослідження, в якій за допомогою мікрочіпів шукають SNP. Якщо вдається виявити варіанти геномів (точніше, сукупність алелів), які значно частіше зустрічаються у людей з даним захворюванням, то кажуть, що такий варіант пов'язаний (або асоційований) з хворобою. На відміну від методів, які перевіряють один або кілька конкретних ділянок геному, повногеномний пошук асоціацій використовує повну послідовність ДНК. Слід зазначити, що цей підхід не виявляє мутацій, що стали причиною захворювання, а тільки виявляє більш-менш значну кореляцію з захворюванням або іншою ознакою. Наприклад, за допомогою GWAS був ідентифікований SNP (заміна G на A) в 5'- нетрансльованій області гена FOXE1, який пов'язаний з підвищеним ризиком раку щитовидної залози.

Ідентифікація генів людини значно збільшує можливості генетичного тестування спадкової схильності і грає важливу роль для медико-генетичного тестування.

4.3 Атлас пухлинних клітин The Cancer Genome Atlas

Вважається, що на сьогоднішній день визначено лише частину впливів, що можуть бути причиною появи раку та можуть використовуватися у практиці як характерні маркери конкретних типів пухлини та/або потенційні молекулярні мішені. Важливу інформацію про біологічну відповідність

молекулярних змін раку можна отримати за допомогою комбінованого аналізу кількох різних типів даних.

Атлас генома раку (TCGA) – це масштабний проект, що розпочався 2006 року, та який підтримується Національним інститутом раку (NCI) та Національним інститутом досліджень геному людини (NHGRI), що є частинами Національного інституту здоров'я, Міністерства охорони здоров’я США та соціальних служб. Цей проект займався картографуванням геномних та епігеномних змін, що спостерігаються у 32 видів раку людини.

1. Його мета – підтримка нових відкриттів шляхом створення каталогу соматичних аберацій, що виникають у різних новоутвореннях, та прискорити темп досліджень, спрямованих на поліпшення діагностики, лікування та профілактики раку; тобто систематизація даних про генетичні мутації, що призводять до виникнення раку. Систематизація проводиться за допомогою секвенування і методів біоінформатики. Інформація, що генерується TCGA, вводиться в бази даних по мірі їх отримання, після їх оцінки за допомогою метрик контролю якості. Станом на 24 липня 2013 року TCGA містила молекулярні структури 7 992 випадків раку, що представлені 27 типами пухлин. До січня 2015 року TCGA генерувала близько 1,7 петабайт даних про близько 11 500 випадків пухлинних захворювань та відповідних нормальних зразків тканин. Деякі файли даних, які не є частиною офіційного набору даних TCGA, можуть розміщуватися на dbGaP. 15 липня 2016 року портал даних був закритий. Зараз дані проекту знаходяться у вільному доступі на порталі Genomics Data Commons.

Успіх аналізу залежить від наявності тканини для аналізу, надійних аналітичних методик/платформ та достовірних даних про результати пацієнтів, які проходили лікування за визначеними та послідовними схемами прийому ліків. Такий всебічний аналіз продемонстровано у проекті TCGA, що показано на рис. 4.1.

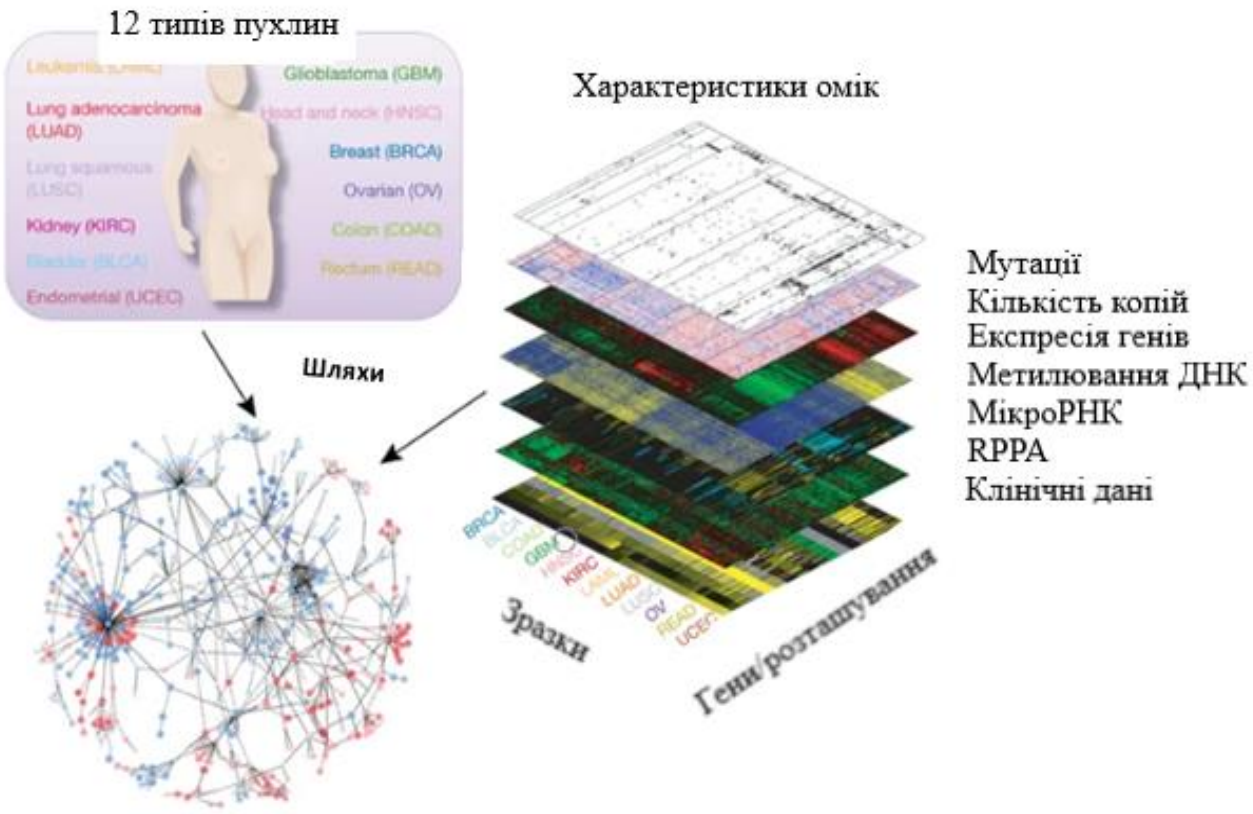


Рисунок 4.1 – Інтегрований набір даних для порівняння декількох типів пухлин

На 2013 рік проект TCGA зібрав дані тисяч пацієнтів з первинними пухлинами, що зустрічаються в різних ділянках тіла, охоплюючи 12 типів пухлин, включаючи мультиформи гліобластоми (ГММ), лімфобластний мієлоїдний лейкоз (LAML), плоскоклітинну карциному голови та шиї (HNSC), аденокарциному легенів (LUAD), рак легень (LUSC), карциному молочної залози (BRCA), ниркову карциному (KIRC), карциному яєчників (OV), карциному сечового міхура (BLCA), аденокарциному товстої кишки (COAD), рак шийки матки та ендометрія (UCEC) та аденокарциному прямої кишки (READ). Шість типів характеристик були виконані, створюючи колекцію даних, в якій елементи пов'язані тим, що для кожного типу використовували однакові вибірки, таким чином, максимізуючи потенціал інтегративного аналізу.

На сайті NCI (National Cancer Institute), що зображений на рис. 4.2, міститься інформація про рак, типи раку, дослідження. Сама платформа TCGA пов'язана з різними платформами та інструментами, що допоможуть отримати інформацію про певний тип раку, типи раку, різні інструменти для обробки даних, тощо.

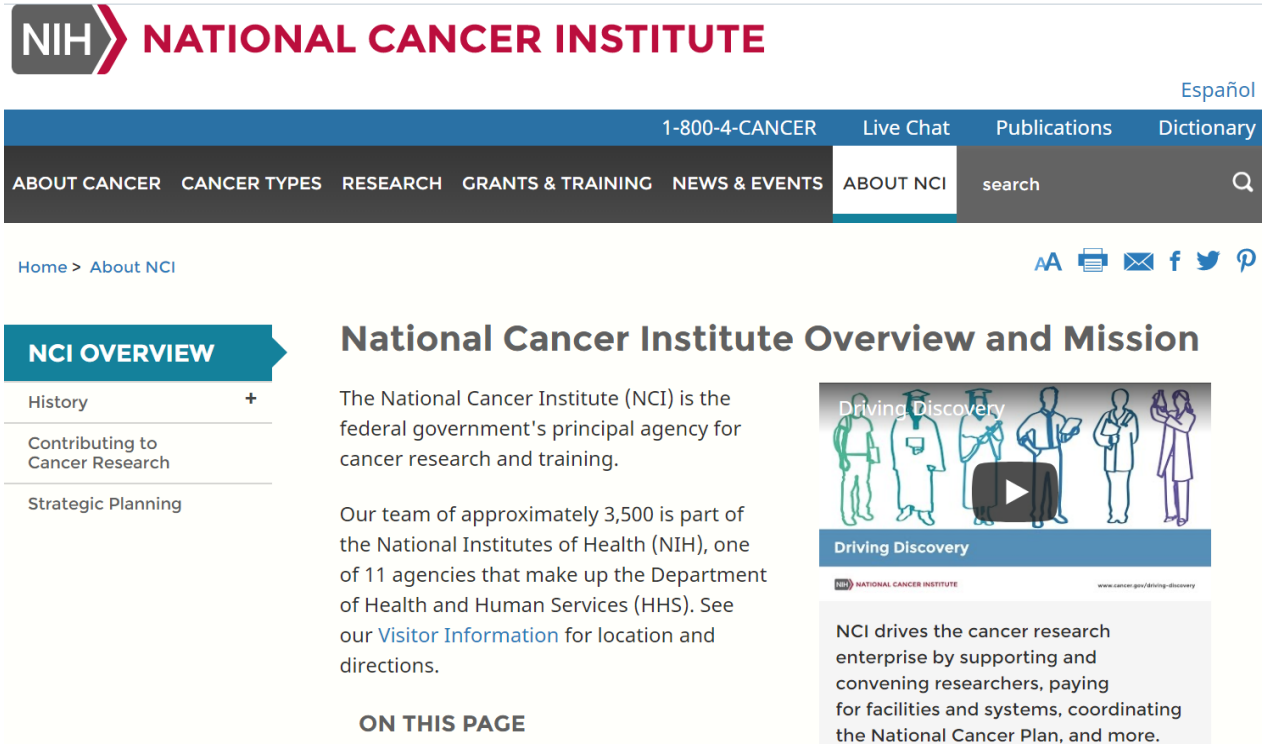


Рисунок 4.2 – Вигляд сайту NCI та інформація про атлас ракового геному

Перед проектом ставляться такі завдання:

1. Зібрати велику кількість високоякісних зразків уражених раком тканин людини, а також їх відповідні зразки нормальних тканин.

2. Послідовно охарактеризувати та проаналізувати кожен зразок.

3. Сприяти співпраці з дослідницькою спільнотою, що досліджує рак.

Дані з TCGA (наприклад, експресія генів, зміна кількості копій та клінічна інформація) доступні через Genomic Data Commons (GDC). Genomic Data Commons Data Portal містить інформацію про дані, їх аналіз, доступ до даних, тощо (рис. 4.3).

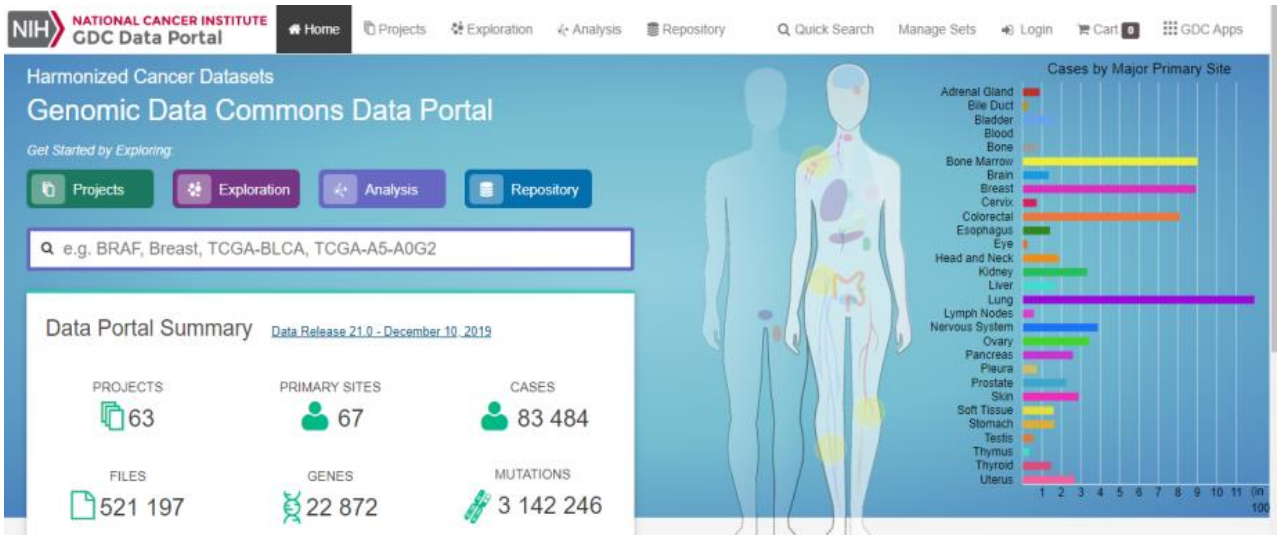


Рисунок 4.3 – Genomic Data Commons Data Portal

Для того, щоб усунути проблему несумісних даних, що були отримані різними дослідниками з використанням різних методів, було укладено контракт з Чиказьким університетом на створення NCI Genomic Data Commons (GDC). Основна мета GDC – це надання даних для дослідницької спільноти з раку. Ця послуга повинна забезпечувати отримання, контроль якості, інтеграцію, зберігання та перерозподіл стандартизованих наборів даних про ракові хвороби. Характерною ознакою високоякісних геномних даних та пов'язаних з ними клінічних анотацій є те, що їх можна комбінувати та видобувати неодноразово, застосовуючи нові алгоритми та методи аналізу. Частиною місії GDC є обслуговування користувачів з широкою різноманітністю біоінформатичних та експертизи даних, надання різних методів перегляду, пошуку та завантаження даних.

Дані TCGA, доступні в GDC, можна класифікувати на три групи типи:

1. Клінічні дані.

2. Дані молекулярного аналізу (геномна характеристика).

3. Метадані аналізу.

Дані TCGA мають три рівні, включаючи рівень 1, рівень 2 та рівень 3. Рівень 1 стосується даних первинної послідовності формату fasta та fastq. Рівень 2 – файл із хорошим порівнянням даних. Рівень 3 визначається як

серія стандартизованих даних. Як правило, якщо ви хочете отримати дані 3 рівня, то слід заповнити заявку на офіційному веб-сайті TCGA. Але більшість звичайних користувачів можуть отримати доступ лише до частини даних 3 рівня. Ці дані наведено в таблиці 4.2.

Таблиця 4.2 – Рівні даних

Рівень даних	Тип даних	Опис
1	Сирі дані	Дані низького рівня, ненормалізовані
2	Оброблені дані	Нормалізовані поодинокі дані
3	Інтерпретовані	Сукупність оброблених даних з одного зразка; пухлина / нормальний зразок.

4.3.1 Рівні доступу та контроль даних в TCGA

Існує два рівні даних, що передбачає відкритий доступ та контрольований доступ.

- Відкриті дані містять інформацію, що не є специфічними для конкретного учасника досліджень. Для отримання таких даних користувач сервісу може не мати відповідних дозволів (сертифікатів). Дані рівня відкритого доступу доступні на порталі даних TCGA.

2. Закриті дані надають інформацію про індивідуальний генотип, тобто унікальний для конкретного учасника досліджень. Доступ до таких даних можливо отримати через авторизований доступ dbGaP Authorized Access.

Типи даних з контрольованим доступом:

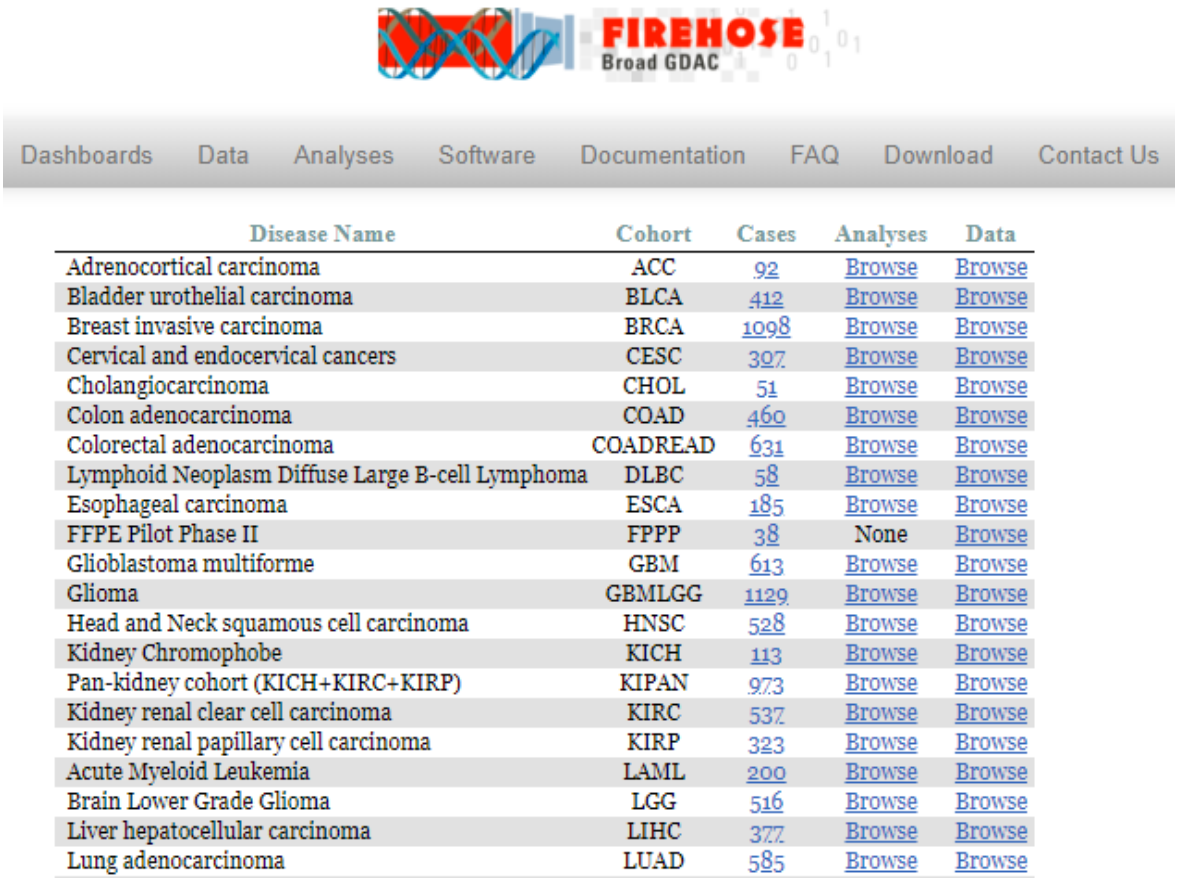
- Індивідуальні дані зародкової лінії (файли SNP.cel),
- Дані про первинну послідовність (файли .bam), які доступні в GDC,

- Клінічні дані,
- Файли масивів Exon.

Існує суворий набір критеріїв, за якими відбувається прийняття досліджень до бази. Інформація щодо зразків пухлин, інформація про зародкові лінії обробляється централізованим сайтом BCR (Biospecimen Core Resource), що послідовно оцінює патологію та дані ДНК та РНК. Усі аналізовані пухлини повинні мати відповідну нормальну пробу від одного пацієнта. У багатьох випадках відповідна норма – це проба крові пацієнта. Опинившись на BCR, всі зразки піддаються контролю якості. Кожен зразок перевіряється, щоб підтвердити діагноз. TCGA вимагає, щоб зразки містили менше 20% некротичної тканини. Як тільки проба проходить огляд на виявлення патологій, виділяють нуклеїнові кислоти і проводять генотипування, щоб кожен зразок пухлини був належним чином пов'язаний з правильною нормальною тканиною. Далі ці проаналізовані зразки проходять процес контролю молекулярної якості і потім розподіляються в TCGA для геномного аналізу.

4.3.2 Інші сервери та аналітичні засоби, пов’язані з TCGA

Сервер Firehose. Дані цього серверу надходять з GDC Data Portal , але поєднують один і той же тип або підтип раку, що дозволяє легко вивантажити всі дані про досліджуване захворювання. Але ці дані не оновлюються в режимі реального часу (рис. 4.4).



Disease Name	Cohort	Cases	Analyses	Data
Adrenocortical carcinoma	ACC	92	Browse	Browse
Bladder urothelial carcinoma	BLCA	412	Browse	Browse
Breast invasive carcinoma	BRCA	1098	Browse	Browse
Cervical and endocervical cancers	CESC	307	Browse	Browse
Cholangiocarcinoma	CHOL	51	Browse	Browse
Colon adenocarcinoma	COAD	460	Browse	Browse
Colorectal adenocarcinoma	COADREAD	631	Browse	Browse
Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	DLBC	58	Browse	Browse
Esophageal carcinoma	ESCA	185	Browse	Browse
FFPE Pilot Phase II	FPPP	38	None	Browse
Glioblastoma multiforme	GBM	613	Browse	Browse
Glioma	GBMLGG	1129	Browse	Browse
Head and Neck squamous cell carcinoma	HNSC	528	Browse	Browse
Kidney Chromophobe	KICH	113	Browse	Browse
Pan-kidney cohort (KICH+KIRC+KIRP)	KIPAN	973	Browse	Browse
Kidney renal clear cell carcinoma	KIRC	537	Browse	Browse
Kidney renal papillary cell carcinoma	KIRP	323	Browse	Browse
Acute Myeloid Leukemia	LAML	200	Browse	Browse
Brain Lower Grade Glioma	LGG	516	Browse	Browse
Liver hepatocellular carcinoma	LIHC	377	Browse	Browse
Lung adenocarcinoma	LUAD	585	Browse	Browse

Рисунок 4.4 – Вигляд серверу Firehose

Портал cBioPortal. cBioPortal є потужним інструментом, розробленим Центром онкологічних захворювань MemorialSloan-Kettering. cBioPortal забезпечує візуалізацію, аналіз та завантаження великомасштабних наборів даних про геноміку раку для Cancer Genomics (рис. 4.5). За допомогою нього можна перевірити кількість певних генів, груп генів, мутацій чи змін при різних ракових захворюваннях та пов'язати їх із певними клінічними ознаками та виживаністю. Можна також використовувати cBioPortal для прогнозування спільної експресії генів або мутацій.

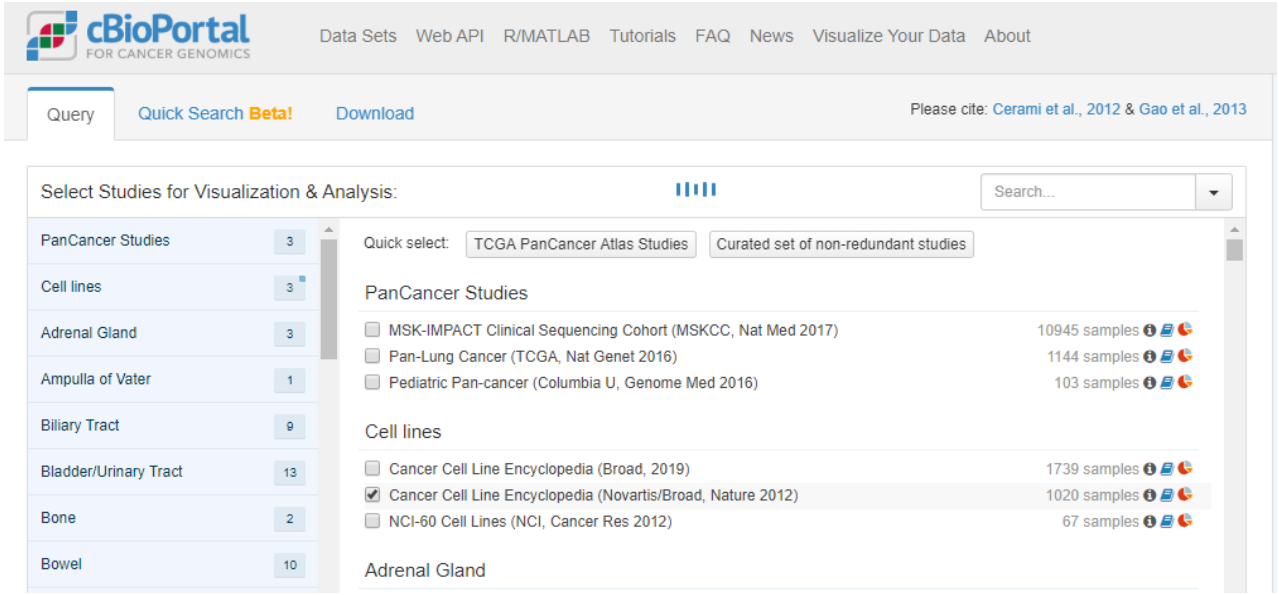


Рисунок 4.6 – Вигляд cBioPortal

Інструмент візуалізації MEXPRESS. MEXPRESS – це інструмент візуалізації даних, призначений для легкої візуалізації експресії генів з TCGA, метилювання ДНК та клінічних даних, а також взаємозв'язків між ними (рис. 4.7).

Рисунок 4.7 – Вигляд інструменту MEXPRESS

Пакет TCGA2STAT. TCGA2STAT дозволяє користувачам легко завантажувати дані TCGA безпосередньо у формат, готовий до статистичного аналізу в середовищі R. Пакет імпортує та обробляє як молекулярні профілі, так і клінічні дані для понад 30 типів раку. Він дозволяє впорядкувати та провести надійний аналіз.

Архів зображень раку. Архів зображень раку (TCIA) – це сервіс, який розміщує великий архів медичних зображень раку, доступних для загального користування (рис. 4.8). Дані організовані у вигляді «Збірників», що містять дані зображень з TCGA, такі як МРТ (магнітно-резонансна томографія), КТ (комп'ютерна томографія) тощо.

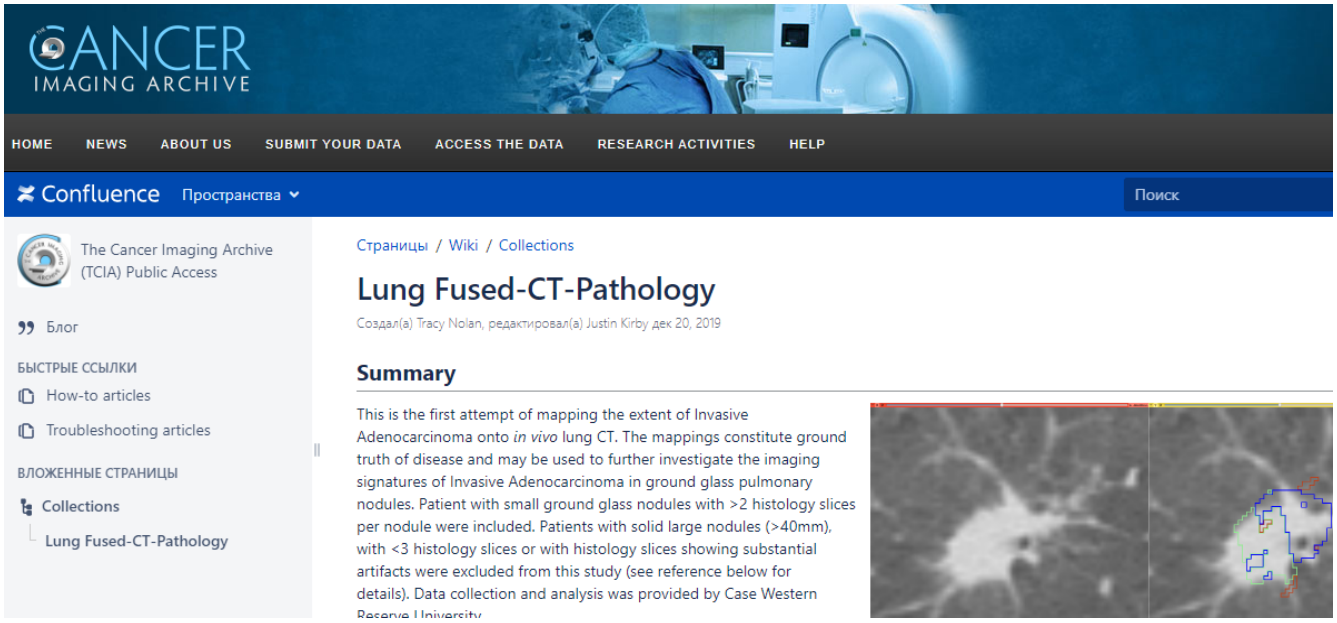


Рисунок 4.8 – Вигляд архіву зображень раку

4.3.3 Атлас пухлинних клітин The Cancer Proteome Atlas

В рамках проекту Атлас геному раку (TCGA) було отримано дані про експресію білка для великої кількості зразків пухлинних та клітинних ліній, за допомогою білкових масивів зворотної фази – Reverse Phase Protein Array (RPPA). RPPA – це кількісна технологія на основі антитіл, яка може оцінити кілька білкових маркерів у багатьох зразках, та аналогічна процедурам

Вестерн-блот. Білки екстрагуються з пухлинної тканини або культивованих клітин, денатуруються SDS (sodium dodecyl sulfate), поміщаються на слайди, а потім додаються антитіла.

Атлас протеома ракових клітин (The Cancer Proteome Atlas) містить найбільший масив загальнодоступних даних про функціональну протеоміку раку з даними ДНК та РНК. Поточний випуск атласу містить 8167 проб пухлин, та в основному складається з наборів зразків пухлинної тканини з TCGA.

Атлас містить два окремих веб-додатки. Перший містить більше 8000 зразків 32 типів раку з Атласу генома раку та інших баз. Другий додаток фокусується на даних ракових клітинних ліній і містить понад 650 зразків 19 ліній.

Цей атлас використовує зовнішні ресурси, до яких відносяться:

TCGA: Атлас геному раку,

HPRD: Довідкова база даних про білок людини,

CCLE: Енциклопедія ракових клітин,

COSMIC: Каталог соматичних мутацій раку,

CTRP: Портал реагування на терапію раку, та інші.

На першій сторінці TCPA є три модулі: підсумок, мій білок, візуалізація, аналіз, як це показано на рисунку 4.9.

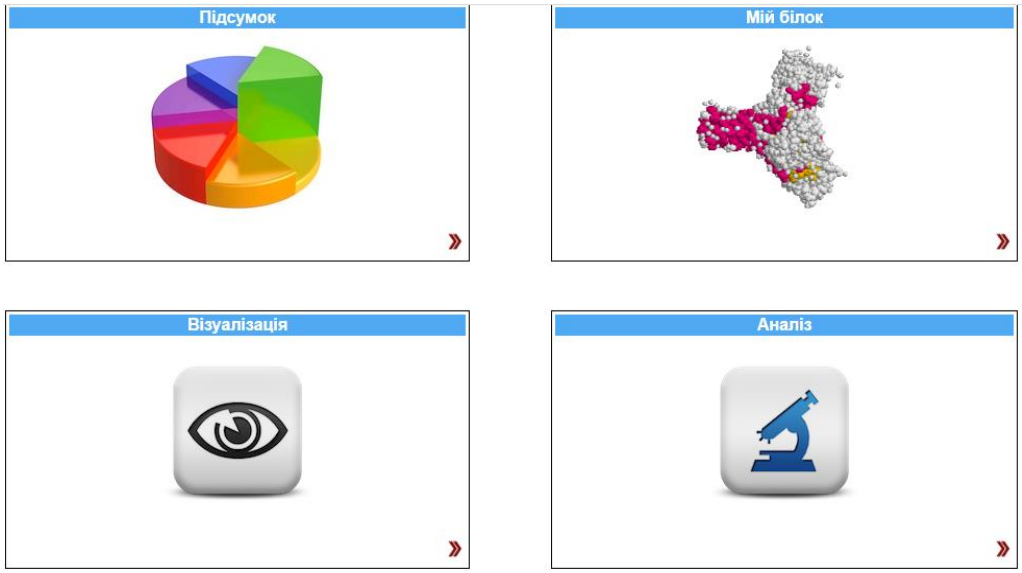


Рисунок 4.9 – Вигляд сайту The Cancer Proteome Atlas

Модуль «Набори даних» («Підсумковий») надає детальну інформацію про зразки та білкові маркери, що знаходяться в TCPA.

Атлас надає можливість окремо завантажити будь-який набір даних для аналізу (рисунок 4.10).

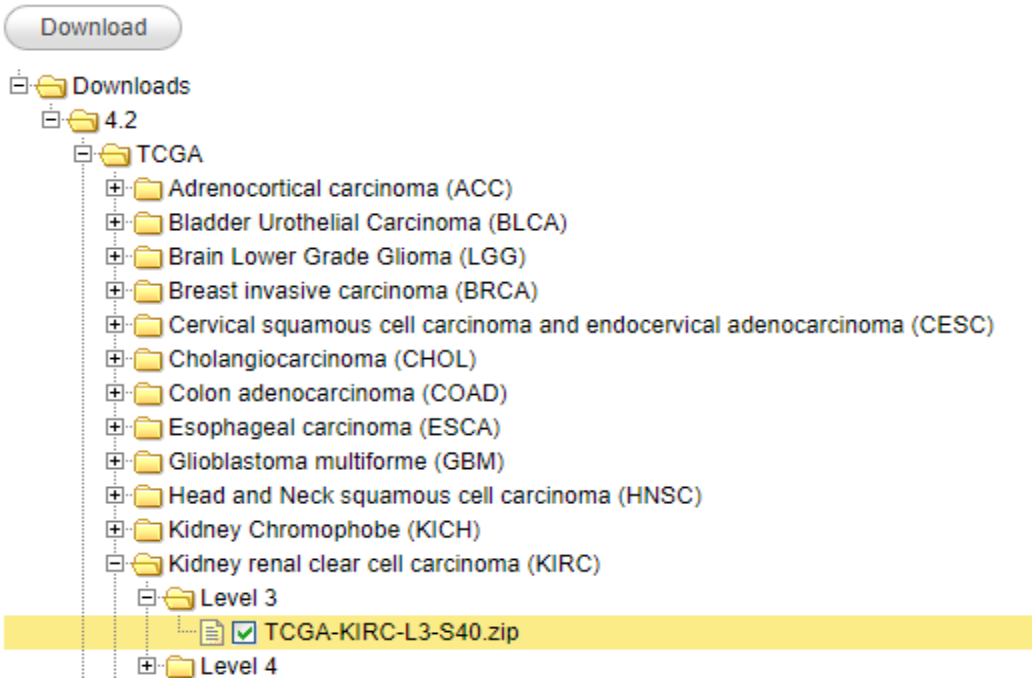


Рисунок 4.10 – Завантаження даних з атласу

Модуль «Мій білок» надає детальну інформацію про кожен білок: назва білка, відповідний символ гена, стан антитіла та джерело антитіла. Користувачі можуть вивчити схему експресії білка. Наприклад на рисунку 4.11 показана експресія HER2 (human epidermal growth factor receptor 2). Для цього атлас використовує інформацію з бази даних GeneCards та інформаційного ресурсу з даними про біомаркери OncoMX. Цей модуль надає детальну інформацію про кожен оцінений маркер.

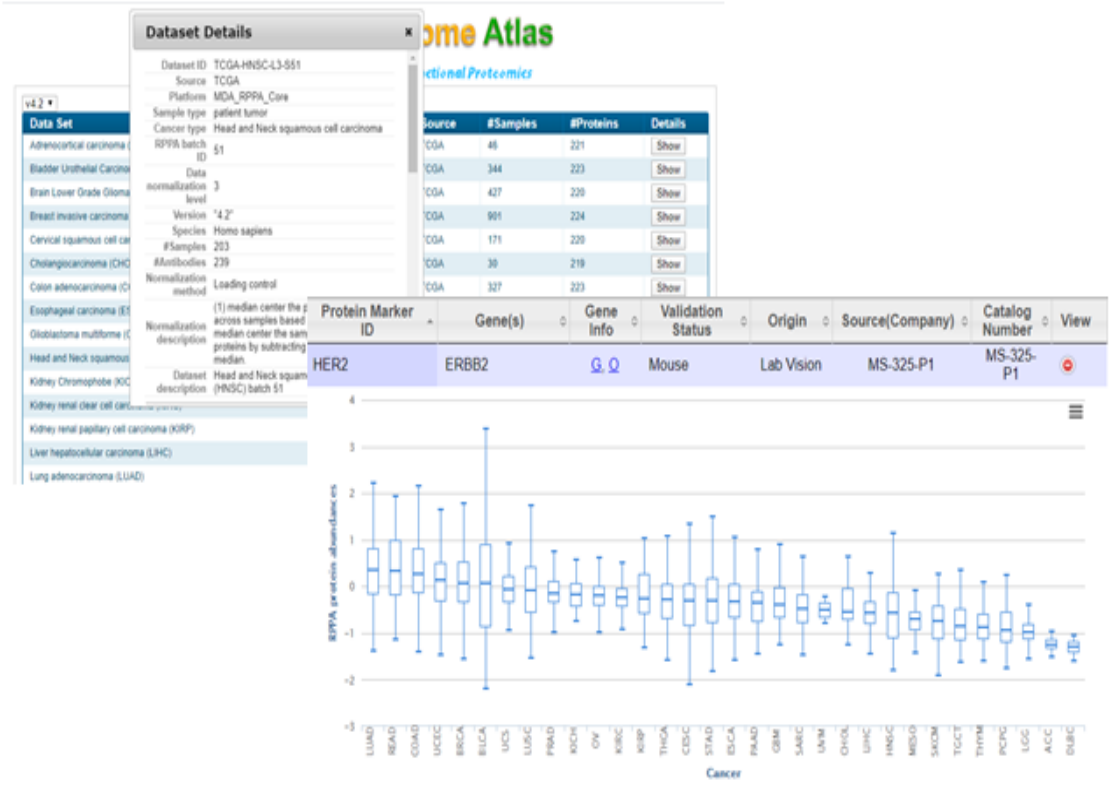


Рисунок 4.11 – Приклад використання модулю «Мій білок»

Модуль «Візуалізація» пропонує два способи вивчення глобальних моделей експресії білка в конкретному наборі даних. Перший – через кластерну теплову карту, яка дозволяє користувачам масштабувати, орієнтувати та вивчати шаблони кластеризації зразків або білків та пов'язувати ці зразки з відповідними біологічними джерелами інформації (рис. 4.12).

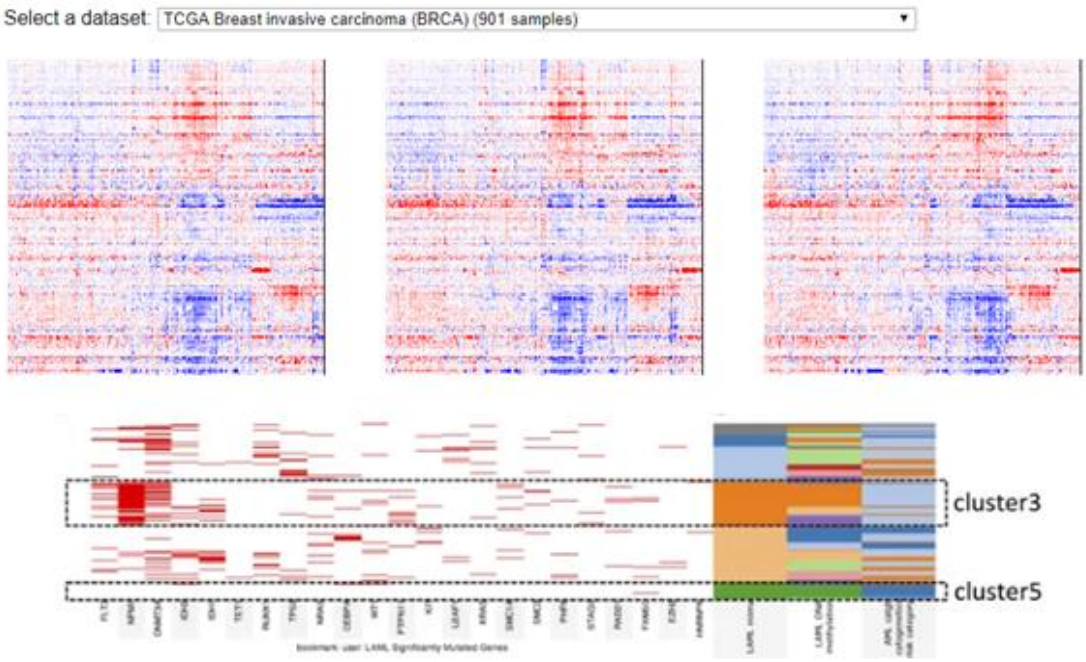


Рисунок 4.12 – Кластерна теплова карта та карта отримана за допомогою UCSC Cancer Genomics Browser

Зразки, що були проаналізовані за допомогою UCSC (University of California Santa Cruz) Cancer Genomics Browser, розташовані рядками, а гени – стовпчиками. Червоний колір вказує на те, що зразки пухлини містять несинонімічні кодуючі мутації у відповідному гені, тоді як білий вказує, що такі мутації не були виявлені. Сильна відповідність спостерігається між кластером 3 – мРНК (помаранчевий), кластером 3 – метилювання ДНК (також помаранчевим) та проміжним цитогенетичним ризиком (світло-блакитний); і між кластером 5 – мРНК (зелений), кластером 5 – метилювання ДНК (також зеленим) та сприятливим цитогенетичним ризиком (темно-синій).

Інша можливість, яку надає модуль «Візуаліазація» – дослідження через мережевий вигляд, який накладає кореляцію між будь-якими двома взаємодіючими партнерами в мережі взаємодії з білками. Мережі показано на рисунку 4.13.

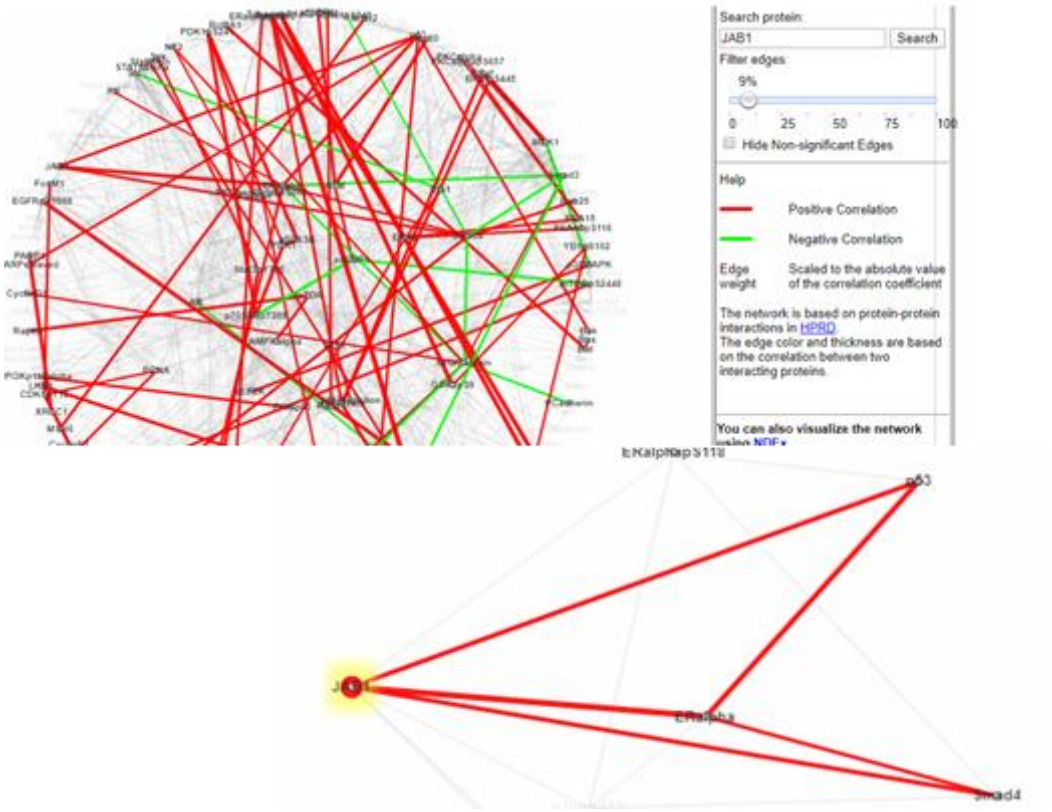


Рисунок 4.13 – Візуалізація мережі взаємодії білків [

Модуль аналізу надає два методи аналізу:

- Індивідуальний аналіз раку,

- Пан-раковий аналіз.

Індивідуальний метод аналізу має три інструмента. Кореляційний аналіз може проводитися між будь-якою парою білків. Користувачі можуть шукати результати за назвою білка, ранговими кореляціями або візуалізувати графік розсіювання. Для подальшої перевірки, чи має цей білок деяку прогностичну чутливість до лікарських засобів, для аналізу того, які препарати співвідносяться з цим білком, можна використовувати «Аналіз на білок-ліки» (рис. 4.14).

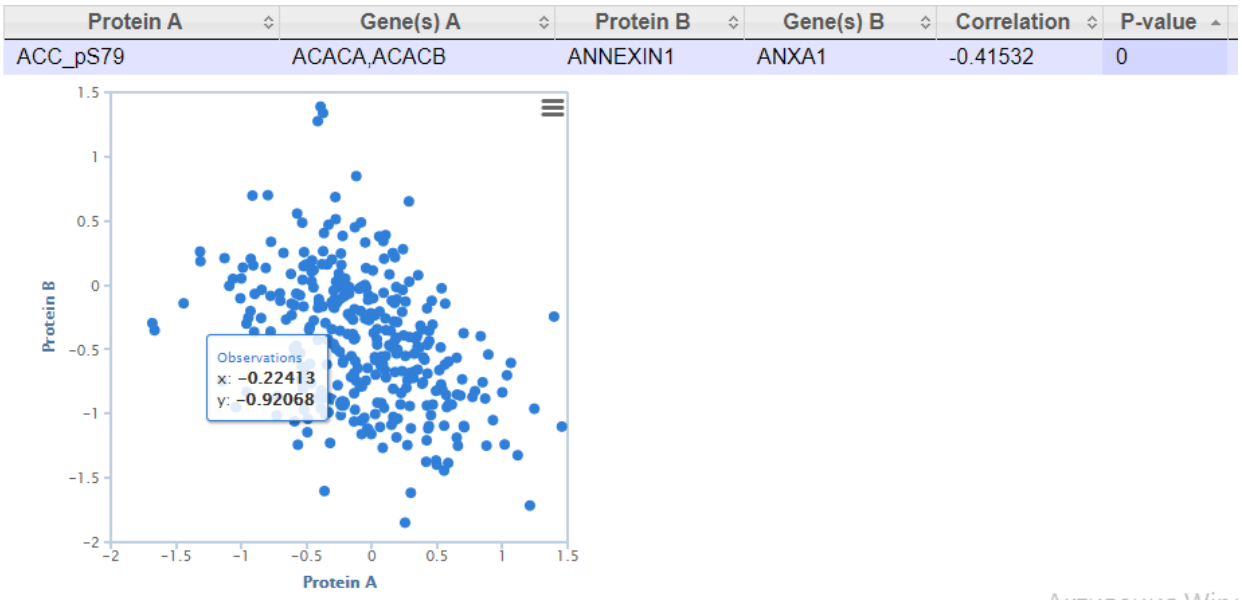


Рисунок 4.14 – Кореляційний аналіз

Для диференціального аналізу можна виділити білкові маркери між двома типами пухлин або підтипами пухлин. З огляду на визначені користувачем групи порівняння, результати відображаються у вигляді таблиці, а для білка, що цікавить, користувачі можуть візуалізувати графіки для порівняння, що продемонстровано на рис. 4.15.

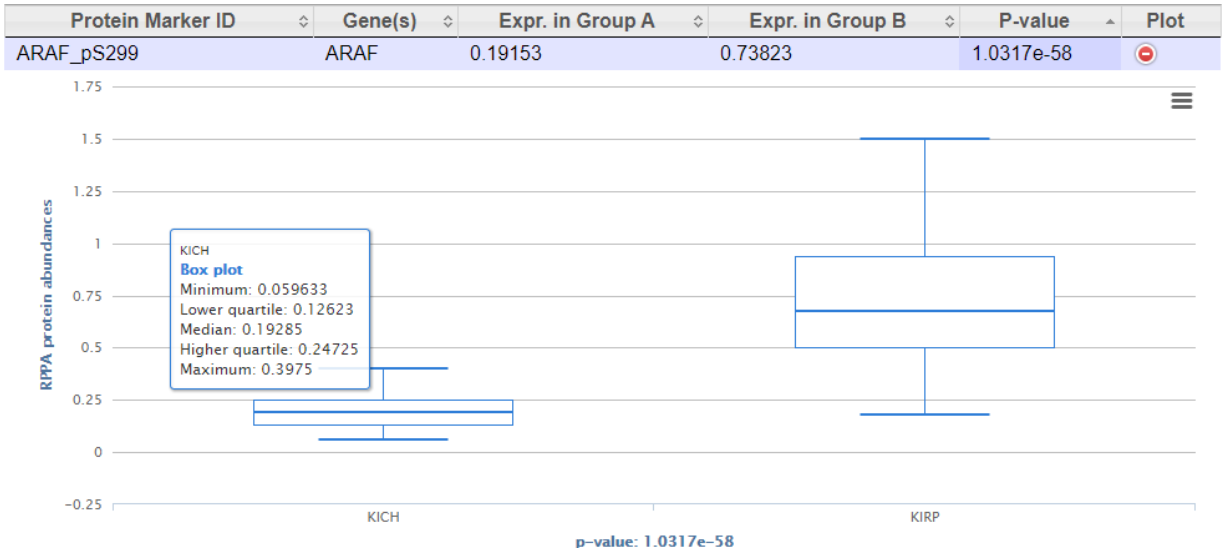


Рисунок 4.15 – Графік порівняння за допомогою диференціального аналізу

Для аналізу виживання можна визначити білкові маркери або події, що суттєво корелюються з виживаністю пацієнта (рис. 4.16). У таблиці показано універсальну модель пропорційної небезпеки Кокса та графік Каплана-Мейєра для кожного білка в наборі даних (наприклад, пацієнти з високою експресією IGFBP2 демонструють кращу виживаність, ніж хворі з низькою експресією).

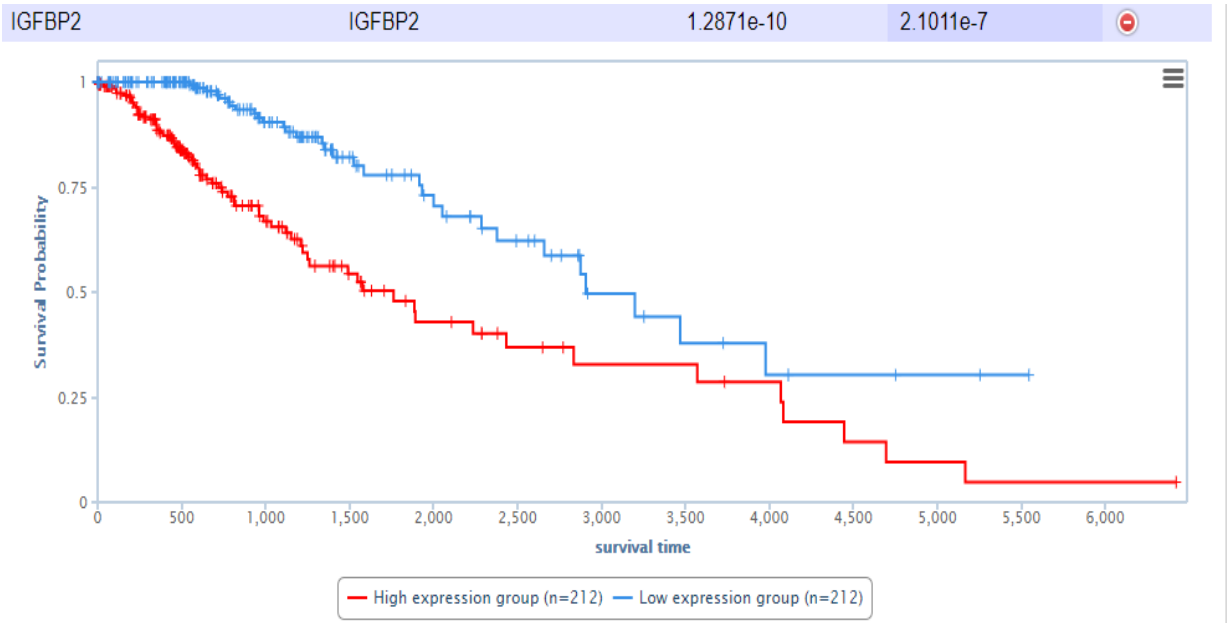


Рисунок 4.16 – Графік виживаності

Модуль «Клітинні лінії» надає можливості для двох аналізів клітинних ліній пухлин. Досліджувані клітинні лінії пов'язані з Енциклопедією ракових клітинних ліній, з якої можна отримати вибрані мутації, транскриптомні профілі та чутливість до конкретних лікарських засобів. Для аналізу лікування препаратами передбачено виявлення впливу лікарських засобів на профілі RPPA.

4.3.4 Перспективи розвитку атласу протеома ракових клітин

У порівнянні з іншими «протеомічними» базами даних, такими як Атлас протеїну людини, перевагою TCPA є наявність кількісних даних експресії

білка у великих групах добре охарактеризованих пухлин TCGA, з пов'язаними аналізами ДНК та РНК.

TCPA дозволяє перевірити висновки даних TCGA RPPA за допомогою незалежних зразків і допоможе користувачам вибрати модельні лінії клітин пухлин для подальшого функціонального дослідження. TCPA доповнює генетичні джерела ракових даних, орієнтованих на дослідження нуклеїнових кислот, такі як CCLE (The *Cancer Cell Line Encyclopedia*), cBioPortal та для онкологічних генологічних центрів Sloan-Kettering Center of Cancer Center, OncoMine та браузеру UCSC Cancer Genomics. TCPA також доповнює інші ресурси, що містять інформацію про білки, такі як Human Protein Reference Database, інструмент пошуку для пошуку взаємодіючих генів/білків та Human Interactome Project.

TCPA – це не лише єдине сховище даних для отримання високоякісних даних RPPA, але і потужна веб-платформа для аналізу цих даних за допомогою зручного для користувача інструменту. В майбутньому планується додавання даних про RPPA раку та інтеграція в цей ресурс інші геномні та клінічні дані. Очікується, що TCPA буде надалі слугувати надзвичайно цінним ресурсом, який буде допомагати дослідникам генерувати перевірені гіпотези, підтверджувати результати, що цікавлять, та врешті сприяти розробці нових методів лікування раку.

Запитання до розділу 4

1. Дати визначення однонуклеотидному поліморфізму.
2. В яких областях наук про життя використовується аналіз SNP?
3. Номенклатура SNP.
4. Назвіть типи однонуклеотидних поліморфізмів кодуючих ділянок.
5. Що таке повногеномний пошук асоціацій?
6. Назвіть приклади досліджень з використанням SNP.
7. Назвіть основні БД SNP.

8. Назвіть основний ресурс для виявлення генетичної мінливості, факторів навколишнього середовища, які впливають на здоров'я.
9. Мета створення Атласу генома раку TCGA?
10. Які завдання ставляться перед проектом TCGA?
11. На які три групи можна класифікувати дані TCGA, доступні в Genomic Data Commons?
12. Який портал забезпечує візуалізацію, аналіз та завантаження великомасштабних наборів даних про геноміку раку для Cancer Genomics?
13. Назвіть сервіс, який розміщує великий архів медичних зображень раку, доступних для загального користування?
14. Яку інформацію містить Атлас протеома ракових клітин?
15. Які зовнішні ресурси використовує Атлас протеома ракових клітин?

Література до розділу 4

1. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. 2002. Interrogating a high-density SNP map for signatures of natural selection. Genome Research, 12(12): 1805–14.

2. Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics. 2005;21(12):2814–2820.

3. Coulet A, Smaïl M Tabbone, Benlian P, Napoli A, Devignes MD. SNP-Ontology for semantic integration of genomic variation data, 14th Annual International Conference on Intelligent Systems for Molecular-Biology - ISMB'06. 2006.

4. Miller RD, Kwok PY. The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. Human Molecular Genetics. 2001;10(20):2195–2198.

5. Su SC, Jay Kuo CC, Chen T. Single nucleotide polymorphism data analysis - state-of-the-art review on this emerging field from a signal processing viewpoint. Signal Processing Magazine, IEEE. 2007;24:75–82].

6. Coulet A, Smaïl-Tabbone M, Benlian P, Napoli A, Devignes MD. SNP-Converter: an ontology-based solution to reconcile heterogeneous SNP descriptions for pharmacogenomic studies. Data Integration in the Life Sciences. 2006;4075:82–93

7. Brookes A.J. The essence of SNPs. / Brookes A.J. // Department of Genetics and Pathology, Biomedical Center, Uppsala University, 751 23. – 1999.

8. Bruce Carlson. SNPs—A Shortcut to Personalized Medicine // Kalorama. – 2008. – №12.

9. Cancer Genome Atlas Research N, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20

10. cBioPortal. URL: http://www.cbioportal.org/

11. Database resources of the National Center for Biotechnology Information (англ.)// Nucleic Acids Res: journal.– 2007.

12. Den Dunnen J.T.. Recommendations for the description of sequence variants (англ.) // Human Genome Variation Society : journal. –2008.

13. Department of Bioinformatics and Computational Biology.
URL: https://bioinformatics.mdanderson. org/public-software/tcpa/

14. E-cadherin expression and prognosis of head and neck squamous cell carcinoma: evidence from 19 published investigations / Xusheng Ren, Jianning Wang, Xuefen Lin, Xuxia Wang. // Onco Targets Ther. – 2016. – №9. – С. 2447–2453.

15. Firehose Server. URL: http://gdac.broadinstitute.org/

16. GeneCards: The Human Gene Database URL: https://www. genecards.org/

17. OncoMX URL: https://www.oncomx.org/

18. Gwas Central. URL: https://www.gwascentral.org/.

19. Horaitis O, Cotton RG. The challenge of documenting mutation across the genome: the human genome variation society approach. Hum Mutat.

20. International HapMap Project URL: http://www.hapmap.org/

21. Jun Li. Explore, Visualize and Analyze Functional Cancer Proteomic Data Using The Cancer Proteome Atlas / Jun Li, Rehan Akbani, Wei Zhao. // Cancer Res.. –

2017. – №77. – С. 51–54.

22. Kaviar URL: https://omictools.com/kaviar-tool

23. Melissa S. Cline. Exploring TCGA Pan-Cancer Data at the UCSC Cancer Genomics Browser URL: https://www.nature.com/articles/srep02652.

24. MEXPRESS. URL: https://mexpress.be/?ref=labworm

25. Miller RD, Kwok PY. The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implications for human history and medicine. Human Molecular Genetics. 2001;10(20):2195–2198.

26. MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs (англ.) // BMC Genomics (англ.)русск. : journal. — 2012.

27. Nachman, Michael W. (2001). Single nucleotide polymorphisms and recombination rate in humans. Trends in genetics 17 (9): 481–485. PMID 11525814. doi:10.1016/S0168-9525(01)02409-X

28. National Institutes of Health The Cancer Genome Atlas (TCGA) URL: https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

28. Nielsen, R. & Signorovitch, J. 2003. Correcting for ascertainment biases when analyzing SNP data: applications to the estimation of linkage disequilibrium. Theoretical Population Biology, 63: 245–55.

29. Sachidanandam, R., Weissman, D., Schmidt, et. al. International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature, 409: 928–933.

30. Shiffman D.,Ellis S.G.,Rowland C.M. et al. Identification of Four Gene Variants Associated with Myocardial Infarction. Am. J. Hum. Genet., 77: 596-605, 2005

31. SNPedia: a wiki supporting personal genome annotation, interpretation and analysis (англ.) // Nucleic Acids Research (англ.)русск. : journal. — 2011.

32. Syvänen, A.C. 2001. Accessing genetic variation genotyping single nucleotide polymorphisms. Nature Reviews Genetics, 2: 930–941

33. Tamura K, Suzuki M, Arakawa H, Tokuyama K, Morikawa A. Linkage and

association studies of STAT6 gene polymorphisms and allergic diseases. International Archives of Allergy and Immunology. 2003;131:33–38.

34. TCGA2STAT. URL: http://www.liuzlab.org/TCGA2STAT/?ref=labworm

35. TCIA. URL: http://www.cancerimagingarchive.net/

36. TCPA: a resource for cancer functional proteomics data / Li J., Lu Y., Akbani R. // Nature Methods. – 2013. – №10. – С. 1046—1047

37. The Cancer Genome Atlas Pan-Cancer analysis project / John N Weinstein, Eric A Collisson, Gordon B Mills та ін.]. // Nature Genetics volume. – 2013. – №45. – С. 1113–1120.

38. The Cancer Genome Atlas Program. URL: https://www.cancer.gov/aboutnci/

39. The Cancer Proteome Atlas. URL: https://tcpaportal.org/tcpa/faq.html

40. The NCBI dbGaP database of genotypes and phenotypes / Matthew D. Mailman, Michael Feolo, Yumi Jin та ін.]. // Nat Genet. – 2007. – №39. – С. 1181–1186.

41. Utkin Lev Vladimirovich, Zhuk Yulia Alexandrovna. A Genome-Wide Association Study using Pairwise Comparison Matrices (англ.) // SPIIRAS Proceedings.– 2016.

42. Wang Z. A Practical Guide to The Cancer Genome Atlas (TCGA) / Wang Z., Jensen M. A.,, Zenklusen J. C.. // Statistical Genomics. – 2016. – №1418. – С. 111–141.

43. Wong, G.K., e.t.c. International Chicken Polymorphism Map Consortium. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. Nature, 432: 717–722.

44. Бородина Т.И. Методы детекции SNP.
URL:http://molbiol.edu.ru/review/0403b.html.

45. Генетичний та асоціативний аналіз однонуклеотидного поліморфізму g.22 g>c у гені катепсину f свиней різних порід. // ВІСНИК Полтавської державної аграрної академії. – 2018. – №4. – С. 137–141.

46. Изучение ассоциации однонуклеотидного полиморфизма rs 231775 гена ctla4 с риском развития бронхиальной астмы. // Красноярский государственный медицинский университет имени профессора В.Ф.Войно-

Ясенецкого. – 2015. – №1. – С. 38–42.

47. Полоников А.В. Полиморфизм генов ферментов биотрансформации ксенобиотиков и его вклад в предрасположенность к мультифакториальным заболеваниям: Диссертация на соискание ученой степени доктора медицинских наук. – М. 2006 г Quackenbush J. Extracting biology from high-dimensional biological data / Quackenbush. // Journal of Experimental Biology. – 2007. – №210. – С. 1507–1517.

48. Пономаренко И.В. Отбор полиморфных локусов для анализа ассоциаций при генетико-эпидемиологических исследованиях // Научный результат. Медицина и фармация. 2018.

49. Селекційні аспекти застосування snp-аналізу у кукуруд : автореф. дис. на здобуття наук. ступеня канд. с.-г. наук : спец. 06.01.05 "селекція і насінництво" / . – Дніпропетровськ, 2014. – 20 с.

50. Состояние всемирных генетических ресурсов животных в сфере продовольствия и сельского хозяйства /ФАО, 2010. ВИЖ РАСХН, 2010. Москва /Перевод с англ. FAO. 2007. The State of the World’s Animal Genetic Resources for Food and Agriculture, edited by Barbara Rischkowsky & Dafydd Pilling. Rome.

51. Фогель Ф., Мотульски А. Генетика человека: в 3-х т. Т.2. Пер. с англ.- М.:Мир. –1990 –378 с.